DOCUMENT RESUME

ED 078 844 LI 004 379

AUTHOR	Strong, Suzanne Marvin
TITLE	An Algorithm for Generating Structural Surrogates of English Text.
INSTITUTION	Ohio State Univ., Columbus. Computer and Information Science Research Center.
SPONS AGENCY	National Science Foundation, Washington, D.C. Office of Science Information Services.
REPORT NO	OSU-CISRC-TR-73-3
PUB DATE	Apr 73
NOTE	148p.;(13 References)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Algorithms; *Computer Programs; *Electronic Data Processing; *Information Processing; Information Systems; *Language Patterns
IDENTIFIERS	Machine Readable Text; *Surrogates

ABSTRACT

An algorithm for generating a structural syntactic surrogate of English text is defined in this paper. The "performance" of the surrogate is judged empirically according to adherence to the following criteria: (1) The surrogate is an organized representation of natural language text; (2) The algorithm which produces the surrogate is equally applicable to any English text; (3) The surrogate may be derived solely by computer processing; (4) The major concepts (or themes) of the text are clearly discernable and/or mechanically derivable from the surrogate; (5) The document may be derived from the surrogate; and (6) The surrogate is feasible to implement. That is, it is not inordinately expensive in computer processing time nor storage, and does not rely upon human processing, preprocessing or interpretation. Possible applications and derivations of the surrogate are explored and discussed. (Author/NH)

ED 078844

LI 004 379

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

(OSU-CISRC-TR-73-3)

AN ALGORITHM FOR GENERATING STRUCTURAL SURROGATES OF ENGLISH TEXT

by

Suzanne Marvin Strong

Work performed under

Grant No. 534.1, National Science Foundation

Computer and Information Science Research Center
The Ohio State University
Columbus, Ohio 43210

April 1973

FILMED FROM BEST AVAILABLE COPY

PREFACE

This work was done in partial fulfillment of the requirements for a master of science degree in Computer and Information Science from The Ohio State University. It was supported in part by Grant No. GN 534.1 from the Office of Science Information Service, National Science Foundation, to the Computer and Information Science Research Center of The Ohio State University.

The Computer and Information Science Research Center of The Ohio State University is an interdisciplinary research organization which consists of the staff, graduate students, and faculty of many University departments and laboratories. This report is based on research accomplished in cooperation with the Department of Computer and Information Science.

The research was administered and monitored by The Ohio State University Research Foundation.

ACKNOWLEDGEMENTS

I would like to thank my thesis adviser, Dr. James E. Rush for his guidance during the research for this thesis and his invaluable assistance in its preparation. I would also like to thank my academic adviser, Dr. James B. Randels for his support and direction during the academic phase of my graduate career.

I am indebted to the following people for their part in the construction of this thesis: to our research group (Dr. James E. Rush, Dr. Harold B. Pepinsky, Dr. Naomi M. Meara, Dr. Carol E. Young, Father John A. Valley and Dr. Walter A. Cook, S.J.) for providing the atmosphere conducive to investigation and cooperation; to Dr. Carol E. Young for her work with syntactic and case grammar analysis precedent to this research; to Dr. Harold B. Pepinsky for his helpful review of the thesis; to Mrs. Mary Kimball for her excellent typing of the manuscript; and to my husband for his encouragement and drafting assistance.

This work has been supported in part by the Mershon Center for Programs of Research and Education in Leadership and Public Policy and in part through a grant (GN 534.1) from the National Science Foundation to the Computer and Information Science Research Center.

TABLE OF CONTENTS

	Page
PREFACE	ii
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	viii
QUOTATION REFERENCES	ix
CHAPTER I. INTRODUCTION	1
1. Motivation	1
2. Design Criteria	3
CHAPTER II. HISTORICAL REVIEW	5
1. Structural Prototypes	5
1.1. Bernier and Heumann	5
1.2. Doyle	6
1.3. Quillian	7
1.4. Fugmann	9
1.5. Shank and Tesler	10
2. Validity of the Syntactic Approach	14
2.1. Klein-Simmons	14
2.2. Thorne, Bratley and Dewar	15
2.3. Clark and Wall	16
2.4. Vigor, Urquhart and Wilkinson	20
2.5. Winograd	21
2.6. Young	22
3. The Place of Case	24
4. Summary	27
CHAPTER III. THE PRESENT RESEARCH	29
1. Phase 1: An Algorithm for Graphic Notation of English Sentences (AGNES)	29
1.1. The Basic Algorithm	29
1.2. Display Extensions	38
1.2.1. Specification of Relator	38

1.2.2. Identification of Implied Relationships 46
1.2.2.1. Stative Verbs 46
1.2.2.2. Pronouns 48
1.2.3. Reductions 49
1.3. Intersentence Relationships 54
2. Phase 2: Experimentation and Testing 59
3. Phase 3: Computer Feasibility: Design Considerations 67
3.1. Assignment and Storage of Edges 68
3.2. Ordering the Construction 71
3.3. Graphic Considerations 74
3.3.1. What to Draw 74
3.3.2. Where to Draw It 74
3.3.3. What to Retrieve 75
4. Phase 4: Investigation into Applications 77
4.1. Indexing 77
4.2. Abstracting 83
4.3. Information Retrieval 83
4.4. Other Disciplines 84
5. Summary 85

CHAPTER IV. RESULTS AND DISCUSSION 86

1. Experimental Results 86
1.1. The Four Major Trends 86
1.2. Observed Trends 93
2. Comparison with Existing Systems 95
3. Comparison with Design Criteria 98
4. A Word on Accuracy 100

CHAPTER V. SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH 102

1. Summary 102
2. The Surrogates in Perspective 103
3. Directions for Future Research 104

REFERENCES 108

DATA REFERENCES 111

APPENDIX A 112
APPENDIX B 135

List of Tables

		Page
Chapter III		
Table 3.1	Storage of Edge Assignments	69
Table 3.2	Order of the Construction	72
Chapter IV		
Table 4.1	Case Assignment Distributions	87
Table 4.2	Revised Analysis	94

List of Figures

		Page
Chapter II		
2.1.	Quillian's "Semantic Network"	8
2.2.	A TOSAR graph	11
2.3.	A Conceptual Dependency Parser Network	12
2.4.	A transformational analysis structure	17
2.5.	Partial versus complete phrase structures	18
2.6.	Phrase structure versus dependency trees	20
2.7.	Salton's semantic graph	25
Chapter III		
3.1.	AGNES graphs of "Age-Old Popcorn"	37
3.2.	The basic verb types	40
3.3.	The case frame matrix	41
3.4.	Extended AGNES graphs of "Age-Old Popcorn"	53
3.5.	A three-dimensional surrogate model	55
3.6.	An AGNES network of "Age-Old Popcorn"	60
3.7.	An AGNES network of "Larry's Trip to Tragedy"	62
3.8.	The word positions returned from screen	76
3.9.	A Predicasts mini-abstract sample	81
3.10.	A structural index entry	82
Chapter V		
5.1.	A schemapiric view of the research	105

Quotation References

Page	
1	H. Borko (ed.), Automated Language Processing, John Wiley and Sons, New York, New York, 1967, 292.
1	H. S. Sharp (ed.), Readings in Information Retrieval, Searecrow Press, New York, New York, 1964, 615.
5	B. M. Stewart, Theory of Numbers, The Macmillan Company, New York, New York, 1966, 183.
5	W. K. Wimsatt, Jr. (ed.), Alexander Pope, Selected Poetry and Prose, Holt, Rinehart and Winston, New York, New York, 1965, 157.
14	R. M. Gorrell and C. Laird, Modern English Handbook, Prentice-Hall, Englewood Cliffs, New Jersey, 1965, 199.
24	A. S. Downer (ed.), William Shakespeare, Holt, Rinehart and Winston, New York, New York, 1965, 208.
29	D. D. Barwick (ed.), Great Words of Our Time, Hallmark Editions, Kansas City, Missouri, 1970, 26.
29	L. Carroll, Alice's Adventures in Wonderland, Airmont Publishing Company, New York, New York, 1965, 120.
59	P. R. Evans (ed.), The Family Treasury of Children's Stories Book II, Doubleday and Co., Garden City, New York, 1956, 6.
67	L. A. Landa (ed.), Gulliver's Travels and Other Writings, Houghton Mifflin Company, Boston, Massachusetts, 1960, 148.
77	D. G. Bobrow, "Syntactic Analysis of English by Computer - A Survey", AFIPS 24 (1963), 385.
86	D. O. Bolander, Quotation Dictionary, Career Institute, Mundelein, Illinois, 1969, 227.
102	J. Hargreaves, Computers and the Changing World, Hutchinson, London, England, 1967, 53.
104	D. D. Barwick, op. cit., 39.

CHAPTER I. INTRODUCTION

> I do not think it would be difficult at all to make the translating machine exercise as good judgment in picking the right word as is exercised by many human translators ...
>
> Vannevar Bush, 1949

> ... because of improvements in electronic computers, the cost of conducting information searches will be substantially reduced in the future and such searches will be conducted as a matter of course. Other predictions concern automatic translation of foreign language documents into English, the abstracting and indexing of technical literature by machine and ... the "automated library."
>
> from an abstract of an article by Benjamin F. Cheydleur, Datamation, 1961

1. Motivation

Despite early hopes for computer aid in the area of natural language processing, very little text published today has more than nominal contact with any computer. Systems such as Libaw's IMbricated PRogram for INformation Transfer (1), a comprehensive information system based on machine-readable text, are largely theoretical. Experimental systems have been espoused in the literature (2), but most are too expensive, too limited, or too dependent upon human initiation and direction to be commercially feasible. By and large, industry has adopted very limited and somewhat unimaginative systems such as the various keyword-in-combination techniques (KWIC and PERMUTERM for example) now used in indexing. If one can assume that those early hopes were not wholly ungrounded and that the ultimate aim of at least some of the research mentioned above was a practical natural language processor, something

has gone amiss. It appears that there may be a missing link in the chain from theory to practice.

It is the purpose of this research to propose that the missing link is in fact a proper representation of the natural language text. It is suggested that this representation should be more highly organized than the text--a "metadocument" so to speak--but that the original text should be derivable from it. Further, the major concepts of the document should be clearly discernible or algorithmically derivable from the representation. And the representation for any text should be completely defined by some programmable algorithm which could be profitably implemented in industry.

The idea is not new. As early as 1961 Doyle (3) urged that such representations, which he called proxies, be developed even though he considered their implementation technologically impossible at that time.

What form might these representations take? It appears obvious that if a more organized representation than the text is required; linear strings will not be sufficient. This thesis describes research on a network form for the representation. Again the idea is not new. Doyle's proxies were networks (3). Others (Quillian (2) for example) have postulated that in general, memory can be modeled by a network organization. It seems plausible then to construct the document representations as networks.

What remains to be determined is the type of organization to be imposed upon the text. Several possibilities exist. Doyle's proxies were determined by statistical correlations between words in the text (3).

Quillian's memory model (2) is semantic in organization, relying on human "understanding" of the text. There are a variety of syntactic analyzers available, including those which employ a phrase structure grammar (4), those which make grammatical class assignments (5) and those which assign deep structures (6). Previous research at Ohio State University (7) has produced an analyzer which assigns both traditional grammar classes and case grammar roles. Since the representation must be entirely programmable and the representation must clearly indicate the major concepts, most semantic models and many syntactic models, which rely upon manually-compiled dictionaries and statistical models which cannot for example, distinguish between "computer technology" and "technology before computers," are eliminated from consideration. The syntactic analyzer (MYRA) developed by Young (8) meets both of the above criteria as well as displaying 93% accuracy and requiring minimal storage and processing time. MYRA is assumed as a basis for this research.

2. Design Criteria

In this paper I shall define an algorithm for generating a structural syntactic surrogate of English text. The "performance" of the surrogate will be judged empirically according to adherence to the following criteria.

1) The surrogate is an organized representation of natural language text.

2) The algorithm which produces the surrogate is equally applicable to any English text.

3) The surrogate may be derived solely by computer processing.

4) The major concepts (or themes) of the text are clearly discernable and/or mechanically derivable from the surrogate.

5) The document may be derived from the surrogate.

6) The surrogate is feasible to implement. That is, it is not inordinately expensive in computer processing time nor storage, and does not rely upon human processing, preprocessing or interpretation.

Possible applications and derivations of the surrogate will be explored and discussed.

CHAPTER II. HISTORICAL REVIEW

> The majority of ideas we deal with were conceived by others, often centuries ago. In a great measure, it is really the intelligence of other people that confronts us in science.
>
> D. Mach

1. Structural Prototypes

> Order is Heaven's first law.
>
> Alexander Pope, Essay on Man IV

It has been mentioned that the idea of graphic representation of text is not new, nor are the concepts of syntactic analysis and case grammar. It is appropriate at this point to discuss some of the research related to the topic of this paper. It is not within the scope of this paper, however, to study in detail research that has preceded the present effort. The ensuing discussion provides neither an in-depth study of selected topics nor an exhaustive list of related research. It is included in the hope that the reader may grasp something of the direction and scope of the work which forms the foundation for this research.

1.1. Bernier and Heumann

One of the early structural models is that of Bernier and Heumann (9) who theorized a "vocabulary ball" consisting of related semantemes. The core of the ball is to be the most abstract concept in a document collection, possibly the semanteme "thing". Succeeding layers are

composed of terms which are more specific and greater in number. The structure could be viewed as a collection of ordered relationships among semantemes. A "first-order" relationship is defined to be the link between two semantemes on adjacent layers (for example "gold" and "metal"). A second-order relationship is defined to be any link which exists between semantemes on the same layer (for example "gold" and "silver"). Bernier and Heumann claimed that these relationships would constitute a comprehensive classification and that any term could be "defined" by the use of all its first-order relationships. The model is essentially static and to be produced manually. It was evidentally never more fully developed, but his concept of dimensionality and structural organization are explored in the present research.

1.2. Doyle

A more complete framework for an information system was developed by Doyle in 1961 (3).

Whereas Bernier proposed a classification based on the semantemes inherent in the documents collected, Doyle classified the documents themselves. He proposed a large associative map, somewhat like a cardiovascular system, where the arteries correspond to relationships involving many documents, arterioles fewer, and so forth. Doyle maintained that flexibility in the "capillary" regions was vital and suggested that these regions could be mechanically determined. He thought that the gross organization however, should be manually determined and fixed. Statistical correlation graphs, called document proxies, were employed to position a document onto the map and to

retrieve documents from it. It is these proxies which bear resemblance to the surrogates of the present research.

In support of the graphic form, Doyle contended that graphs are easily evaluated at a glance, that differences among documents may be made conspicuous and that graphs are subject to dynamic control. The present research proposes to demonstrate these points.

1.3. Quillian

Another view of networks and meaning was suggested by Quillian in his development of a model of memory to be used in a language comprehension program (2). Quillian's "semantic networks" maintain the hierarchy displayed by Bernier and Doyle, but allow for predication or modification of a concept within the structure. All factual information is encoded either as a "unit" or a "property". A unit represents some thing, idea, event, etc. and a property represents any sort of predication. Each unit must contain a link to its superset, and may contain links to properties. Each property must contain links to an attribute and a value and may contain links to other properties. The attribute-value pair may represent the traditional category-value pair (for example color-white) or any verb-object or preposition-object pair. Figure 2.1 may help to explain the model.

The unit (in brackets) represents the concept "client". It links to its superset "person" and its property whose attribute is "employ" and value is "professional". This property links to another whose attribute is "by" and whose value is "client". In English, a client is a person who employs a professional. Quillian employed this memory

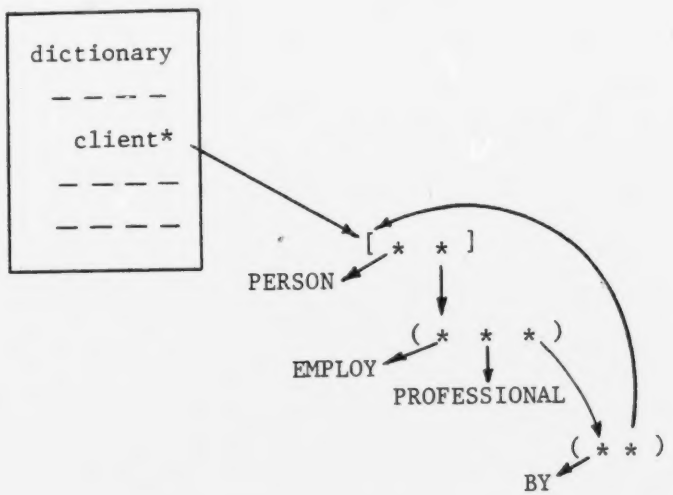


Figure 2.1 A representation of Quillian's "semantic network" organization for "client" as a "person who employs a professional".

model in his "Teachable Language Comprehender" (TLC). In theory, the TLC can be taught to alter its memory structure with "teacher guidance". Initially, however, it was "spoon-fed" a semantic structure.

Quillian's work, though primarily a semantic model, makes use of certain syntactic relationships (such as predicate-object) within the structural framework. This extension over the name hierarchy proposed by Bernier and by Doyle is central to the present research.

1.4. Fugmann

A somewhat different approach to structural organization is displayed in Fugmann's "Topological Method for the Representation of Synthetic and Analytical Relations of Concepts" (10). TOSAR is an attempt to expedite literature searches by a predictable indexing scheme. Unlike Bernier, Doyle and Quillian, Fugmann makes no attempt to structure the document collection. Rather each document or query is represented as an independent network. Organization is not a concept hierarchy, but rather a type of time-line. For each document or inquiry, a graph is drawn manually. The graph is then coded and stored (or mechanically matched) with the graphs which represent the document collection.

The relations displayed by the graph are chemical in nature, but Fugmann asserts that "a method of representing relations between concepts precisely and clearly smooths the road to a consistant analytical treatment even of concepts that are not concerned with structural chemistry" (11).

Each process that is carried out is represented by a series of levels. The concepts before the process are arranged on one level and the concepts after the process are displayed at one point on a lower level. For example, the graph in Figure 2.2 represents the following process.

A, B and C undergo a reaction a which leads to new substances D and E while C is still present. C, D and E are combined with F under b; G is formed from D and F (C and E still present). C is removed by g; E,G are combined with H under d to produce E,H with G separating. Like Fugmann, the graphs in the present research will be representative of the individual document. They will, however, be organized syntactically and produced mechanically.

1.5. Shank and Tesler

Shank and Tesler have gone one step farther and based their structural representation on the sentence (12). Their Conceptual Dependency Parser operates on one word of an input sentence at a time, checking potential links to other words in the sentence with its knowledge of the world and past experience. A linked network is displayed upon the completion of each sentence. A typical network is shown in Figure 2.3. The parser employs a five-step process to construct the network: a dictionary look-up, application of realization rules, an idiom check, rewrite procedures and a semantics check. The dictionary consists of a list of "senses" each composed of a "conceptual category" (such as actor, action, location, etc.) and an "interpretation" (such as: fly--an insect). Guesses as to which sense

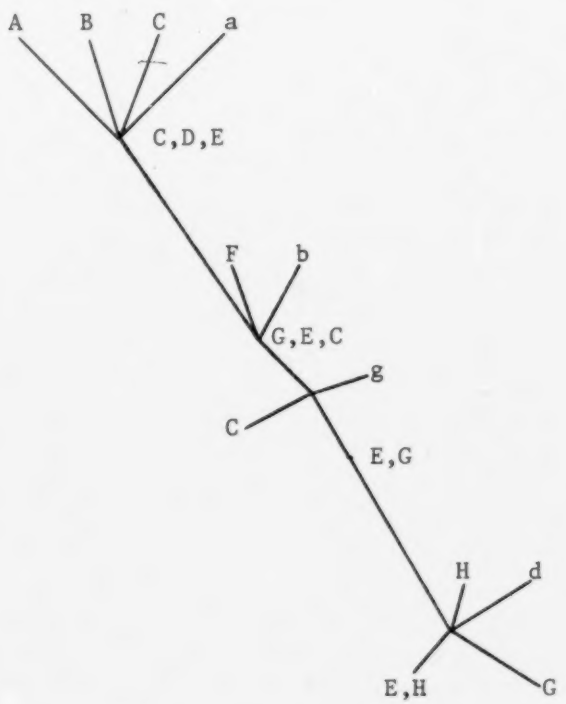


Figure 2.2 A TOSAR graph for a hypothetical chemical reaction.

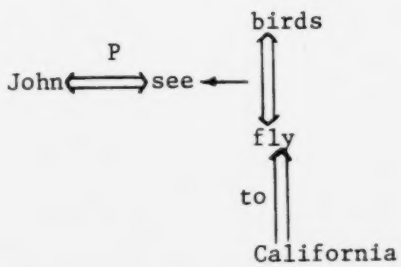


Figure 2.3 Example of a network produced by Shank and Tesler's "Conceptual Dependency Parser".

applies are stacked for testing against the realization rules and conceptual semantics. The rules define how the word might be connected to previous and future words in the sentence to form the structure. These rules prohibit such constructions as the now famous "ideas sleep" on the basis that such a connection has never been made (in the parser's experience). The semantics check is employed to choose between two feasible interpretations. For example, the parser would attempt to construct

"John saw Texas flying to California"

in the same way it handled

"John saw the birds flying to California"

until the semantic check disallowed the construction

"Texas $\Longleftrightarrow$ fly"

Although Shank and Tesler have stressed that the parser is a conceptual rather than a syntactic model, the categories they propose align closely (in English) with the traditional grammar classes and case grammar roles. It was early recognized that it would be advantageous if dictionary look-ups could be replaced by "a computational procedure independent of vocabulary size and number of rules in the grammar" (13). Since grammatical class assignments can be made without reference to the large dictionaries and experience lists Shank and Tesler used, the question arises, "Could a similar network be constructed

relying solely on the grammatical class assignments?"

A brief look at what has been done in syntactic analysis is in order.

2. Validity of the Syntactic Approach

> The congruent and harmonious fitting of parts in a sentence hath almost the force of knitting and connexion: As in stones well squar'd, which will rise strong a great way without mortar.
>
> Ben Jonson

2.1. Klein-Simmons

Klein and Simmons have developed a "computational grammar coder" using a dictionary of fewer than 2000 entries (14). The coder recognizes thirty grammatical classes including the traditional adjective, adverb, noun and the more precise categories such as relative pronoun. The coder assigns each word in the sentence to a class depending on its form, function and/or distribution. The dictionary consists of function words such as articles and prepositions and an exception list of, for example, non-"ly" adverbs. Each word is subjected to several tests including dictionary look-up, capitalization tests, suffix tests and a context frame test. At each step all possible class codes are recorded and at the completion of the tests for each word, the set of all common codes is assigned to the word. The results are "usually unique" (15). If the set is empty, the word is coded "NONE". Klein and Simmons claim 90% accuracy, but there are some apparent inconsistancies. For example, in the sentence "He chose the beautiful" the word "beautiful" is classed as an adjective, but in the sentence "He chose the red",

"red" is classified a noun.

The Klein-Simmons program is an important step in computerized syntactic analysis because the computational approach eliminates large and subject-specific dictionaries.

The Klein-Simmons program is modeled after what may be called a traditional class grammar. That is, the output consists essentially of parts of speech or subdivisions of the same. Other approaches to syntactic analysis have been proposed and subsequently programmed. One such approach is the transformational grammar described by Chomsky (16). A version of a transformational grammar has been programmed by Thorne, Bratley and Dewar of the University of Edinburgh (6).

2.2. Thorne, Bratley and Dewar

A transformational grammar consists of a base component and a transformational component. The base specifies a set of strings which correspond to simple sentences. The transformations combine these into more complex sentences. Thorne, Bratley and Dewar differ from most transformationalists in that they have chosen a regular grammar rather than a context-free phrase-structure grammar as their base component.

The analyzer is provided with a "closed-class" dictionary of function words and suffixes and a finite state network of the grammar (see also Woods (17)). Each dictionary entry may contain several codings of a single word. Each coding consists of a category name and a set of syntactic features (number, tense, etc.). The analysis produces a diverging tree based on predictions made at each step or node of construction. There are four classes of predictions: immediate, which

relates two existing nodes; lexical, which creates a new node on the basis of dictionary information; transformational, which creates a new tree or root node; and terminal, which may reactivate an inactive node. When the end of the sentence is reached, all possible paths recorded on the analysis structure are printed. Figure 2.4 illustrates an analysis structure which has two paths.

Because Thorne, Bratley and Dewar intended the processor to be in itself a psycholinguistic experiment, they imposed several constraints which they felt are constraints in human language processing. They required single pass predictive techniques and simultaneous analysis of surface (or syntactic) and deep (or semantic) structure. For the purposes of this research such restrictions are artificial and the analysis is more complex than required, but the Thorne, et al., study demonstrates that successful automatic analysis using even a complex grammar is possible.

2.3. Clark and Wall

Another approach, taken by Clark and Wall (4), demonstrates that a relatively simple analysis may yield substantial syntactic information. Clark and Wall developed a limited phrase-structure parser for use in mechanized indexing. They adopted a "computational" dictionary similar to the one developed by Klein and Simmons. The grammar they employ identifies only phrases and clauses and no attempt is made to mark relations between phrases. The difference between a "complete" phrase structure and the Clark and Wall model is illustrated in Figure 2.5.

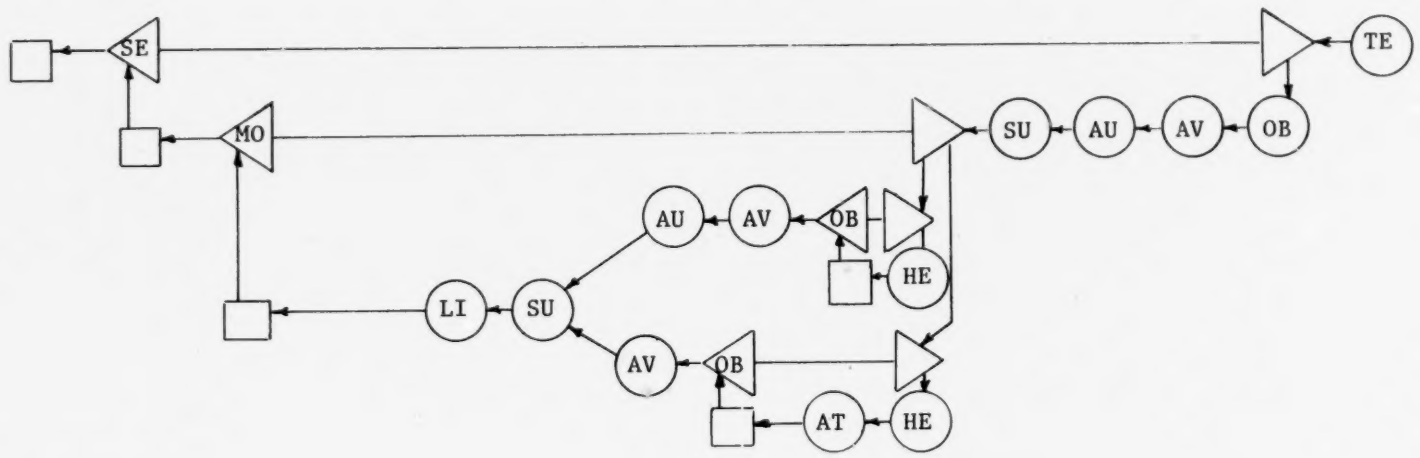
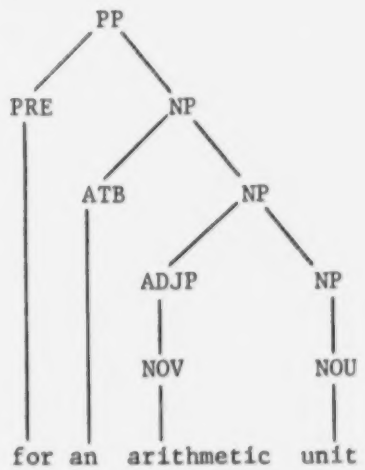


Figure 2.4 A two-path analysis structure produced by the Thorne, Bratley and Dewar grammar.

a.



b.

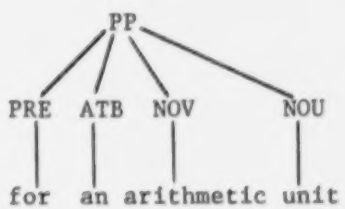


Figure 2.5 a. A tree structure representation of a complete phrase-structure grammar.

b. A tree structure representation of the Clark and Wall modified phrase-structure grammar.

The Clark-Wall algorithm is a set of procedures for assigning an allowable syntactic structure to an input string. The first pass incorporates a dictionary look-up and phrase boundary placement. Pass two establishes clause boundaries and tests well-formedness. Clark and Wall report an average accuracy of 91%. Though encouraged with these results, Clark and Wall are careful to emphasize that the parser is of a limited nature and is "scored" solely on correct identification of phrases.

A report of their research is included here as somewhat of a foil to the Thorne, Bratley and Dewar study. That is, to emphasize that although complicated models of grammar can be successfully programmed, simple models can provide much syntactic information economically.

2.4. Vigor, Urquhart and Wilkinson

Another theory of grammar is employed by Vigor, Urquhart and Wilkinson in the parsing algorithm for their Parsing Recognizer Outputting Sentences in English (PROSE) (18). Their analysis is based upon a dependency grammar analogous to the one described by Hays (19). Figure 2.6 contrasts a phrase structure (such as used by Clark and Wall or Thorne, Bratley and Dewar) with a dependency analysis.

A phrase structure postulates a hierarchy of phrases and sub-phrases comprising the nodes of a tree. In a dependency grammar, the words of the text are the internal nodes of the tree. Vigor, Urquhart and Wilkinson have identified two measures of dependencies which they call binding and determinacy. Binding is a function of physical position. Determinacy is a function of possible contexts. The two values for

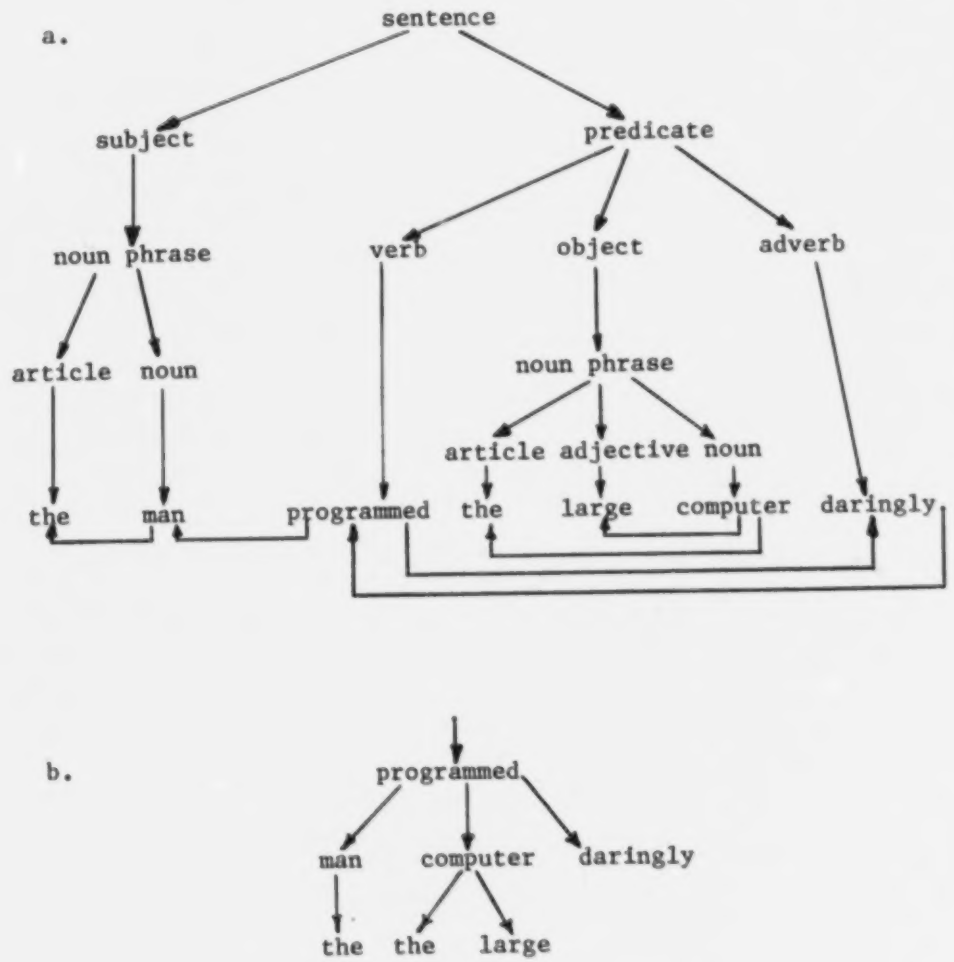


Figure 2.6 a. A phrase-structure tree of a sentence with dependency relationships between sentences marked.

b. A diagram of a dependency-analysis tree structure as is employed in the PROSE project.

binding are bound or loose; the two values for determinacy are determinate or recursive. The relationship of an article and its noun is bound determinate. The relationship of a conjunction to the words it relates is loose recursive.

Vigor, Urquhart and Wilkinson describe the parsing process as setting up a dependency tree by "plugging in plugs". For example, "the" has a "plug" which can only be satisfied by a noun "outlet". "Sits" has a required "plug" for a singular noun or pronoun and an optional "plug" for a positional preposition.

The reader will notice a resemblance between the graphic representation of PROSE and the ones presented in this research. In particular, the centrality of verbs and the dependency "plugs" are essential both to the graphic output and to the syntactic analysis which precedes it.

2.5. Winograd

Still another model of grammar has been used as a basis for Winograd's "robot" language-understanding system (20). The parsing algorithm of the system is based on Halliday's systemic grammar (21). The language is divided into units ranked clause, group and word. Clauses are subdivided into primary or secondary; groups are subdivided into noun groups, adjective groups, etc. Each unit is assigned a set of features (time, number, etc).

The parse proceeds as follows. A clause structure is assigned based on the lexical information of the first word. Succeeding words are matched to the structure to fill possible group structures. Complete groups are stored in a pushdown list until a word is encountered which

does not "fit". A new clause structure is then selected on the basis of the contents of the pushdown list and the current word. Several aspects of Winograd's model are interesting with regard to the present research, but it should be noted that his program operates on contextually limited text with a large dictionary.

2.6. Young

The approach to syntactic analysis which will be assumed for the present research was programmed by Young (8) based on the Fries model of structural classes (22). Fries' model consists of five classes which correspond roughly to noun, verb, adjective, adverb and other. The latter class, which he calls function words, is subdivided into 15 groups. The importance of the Fries model for this research lies in his approach to classification. All words are classed by structure, rather than by meaning. The Young analyzer (MYRA) scans a sentence for function words, checking against a dictionary of about 400 words. On the second pass, all blank elements are classified according to a set of rules. The third pass tests for the occurrence of a verb. If no verb is found, MYRA looks for possible verb slots and reassigns positions accordingly. For example, the first pass over the sentence

"The boy hit the ball."

yields

DTR XXX XXX DTR XXX EOS

(determiner - - determiner - - end-of-sentence)

The second pass encounters two slots after a determiner and assumes

ADJ NON. Thus Pass 2 yields

The boy hit the ball.

DTR ADJ NON DTR NON EOS

(determiner adjective noun determiner noun end-of-sentence)

Pass 3 recognizes that no verb has been assigned and searches for a possible reassignment yielding

The boy hit the ball.

DTR NON VRB DTR NON EOS

(determiner noun verb determiner noun end-of-sentence)

Young claims 93% average accuracy for MYRA. Other advantages of the program include the size of the dictionary, the absence of suffix checks, and consistancy of assignment according to function. In the previous example, "He chose the red" and "He chose the beautiful", MYRA treats both "red" and "beautiful" as nouns because they are single elements following determiners. Further, the analysis tends to isolate errors since no trees are generated, and it has been applied with consistent accuracy to several types of English text.

It is apparent that syntactic procedures yield sufficiently accurate results to produce a parsing structure. The question is whether or not such a structure "tells" us anything about the text. Does it structure the text according to "meaning" or according to arbitrary syntactic constraints. There are many (12, 20) who feel that work in syntax void of semantic considerations is fruitless. They have turned to large dictionaries and semantic checking to resolve ambiguities. For the

purpose of this research however, it is not important whether the program understands what each sentence "means" but rather that it must be able to ascertain what the text is "about". What is needed are some types of "semantic clues" which may be developed from the "syntactic cues" of the text. This research well assume as a basis, Young's case grammar analysis.

3. The Place of Case

> Polonius: What do you read, my lord?
> Hamlet: Words, words, words.
> Polonius: What is the matter, my lord?
> Hamlet: Between who?
> Polonius: I mean the matter that you read, my lord.
>
> Wm. Shakespeare, Hamlet

There is evidence to support the belief that case grammar roles are indeed the semantic clues one might need in language processing. Winograd's "features" parallel case grammar roles and in fact Halliday's systemic grammar may be viewed as a case grammar system (23). His features are assigned solely on the basis of lexical entries, however. Salton alludes to a human-based "semantic graph" (Figure 2.7) which has links labeled "location", "possession", "identity", etc. (24). Salton proposes that "indicated" relations such as location could be automatically determined.

The theory of case grammar was developed by Fillmore and first described in 1968 (25). He defined the sentence to be a modality plus a proposition. The modality constituent embodies the whole sentence and includes such elements as negation, time, mood, etc. The propositional constituent is a set of noun-verb relationships, called cases. The

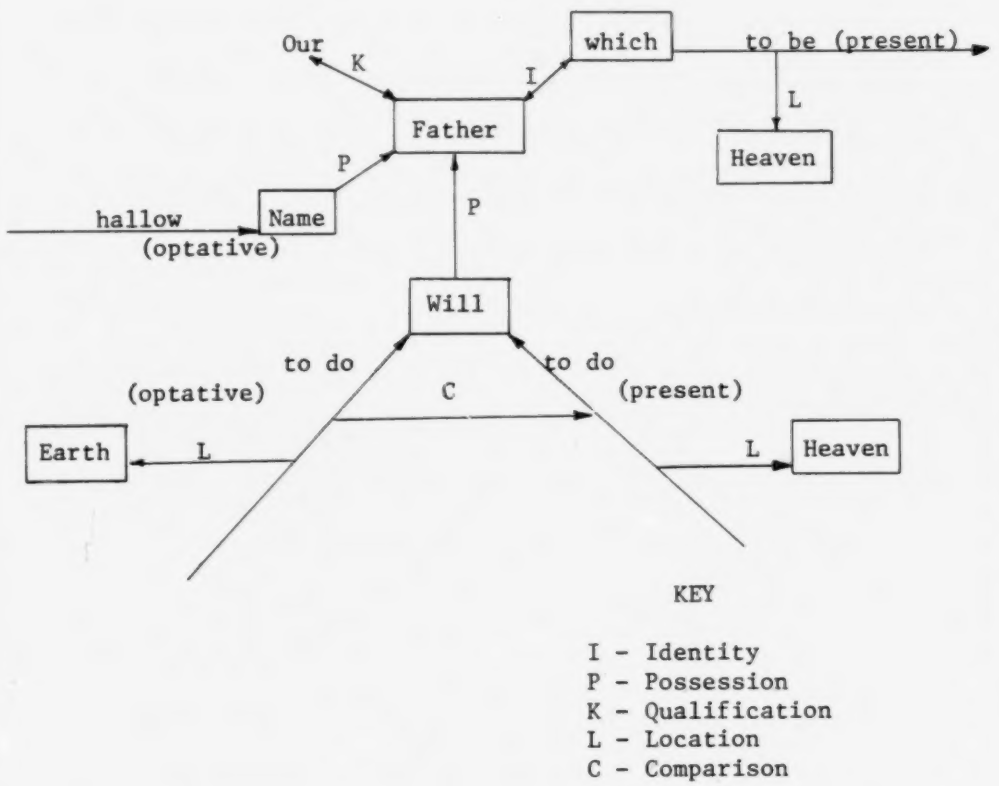


Figure 2.7 Salton's manually produced semantic graph for the text, "Our Father, which art in Heaven, hallowed be thy Name. Thy Kingdom come, thy Will be done on Earth as it is in Heaven."

cases are labeled agentive, instrument, dative, factitive, locative and objective, depending upon the role of the noun in the relationship. It is the noun which in the Fillmore analysis, determines the case. Verbs are classified according to the cases which accept them.

A similar model of semantic structure is proposed by Chafe (26). A sentence in his analysis consists of one "predicative element" and optionally a set of "nominal elements". Chafe takes the position that "it is the verb which dictates the presence and character of the noun" (27) and not the noun which controls the verb. Chafe then defines four "selectional units": state, process, action and ambient. Verbs are classified according to these units and it is these units which "select" the relationship of the noun to the verb. These relationships are the propositional cases Fillmore proposed. Chafe labels them experiencer, agent, benefactive, patient, instrument, locative, and so forth.

Another version of a case grammar was published by John Anderson (28). The labels he adopts for the case roles are different from those of Fillmore and Chafe, but his analysis is similar to theirs. Like Chafe he stresses the centrality of the verb. His study is important to this research because he suggests certain syntactic tests for distinguishing between cases (for example a stative/non-stative test).

The Young implementation of a case grammar analysis (7) recognizes the "essential cases" of agent, experiencer, beneficiary and object and the "peripheral cases" of locative, time, manner, comitative, cause and purpose. The analysis is a three-phase process. Phase 1 identifies the verb class by dictionary look-up or default. Essential cases are

assigned based on the verb class case frame or secondary relators. Phase 2 assigns the time case based on dictionary recognition and phase 3 assigns the remaining peripheral cases based on prepositions and other syntactic and lexical cues.

4. Summary

In summary, the direction of the present research is similar to that proposed by Christine Montgomery (29). She has taken as her definition of content analysis, "a process which ideally involves the identification of the concepts contained in the information records and requirements [queries], and the determination of the relations linking these concepts" (30). She labels the identification a semantic analysis and the determination of the relationship a syntactic analysis. Further, she recognizes that the most "solid achievements" in computational linguistics have been in syntax, but that there appears to be an emergence of some concensus of opinion about semantic fundamentals such as the centrality of the predicate.

She suggests that Fillmore's case grammar provides a "linguistically-based formalism" for representing content in terms of relationships, although she suggests a lexical rather than computational approach. Finally, she suggests a network data structure for storage of these content representations.

Montgomery's proposals for research differ from the present study in that she is concerned with "logical semantic" relations--those which rely upon prior experience rather than contextual inference. Therefore

she is led to propose an elaborate "encyclopedia" similar to Quillian's "semantic networks".

Specifically, the present research builds upon the notion of a structural representation of "concept". However, unlike Bernier and Fugmann, this researcher insists upon automatic production of the structures. Like Quillian's model, syntactic relations involving a predicate will be central to the study. In fact, the basis for the structures will be syntactic rather than statistical as in Doyle's model, or "logically semantic" as in Quillian's, Shank and Tesler's and Montgomery's models.

A computational approach to syntactic analysis similar to the Klein-Simmons approach will be taken. The grammar will be the Young adaptation of the Fries structural grammar because other grammars which have been explored are unsatisfactory for the research purposes. Transformational analyses such as that developed by Thorne, Bratley and Dewar provide more data than is needed. Traditional grammars, such as the one used in the Klein-Simmons analysis may be inconsistant. The systemic grammar implemented by Winograd requires a large dictionary, as does the dependency grammar programmed by the PROSE group. Phrase structure grammars such as produced by Clark and Wall do not provide sufficient relational information.

Finally, the present research will rely on case roles as assigned by the Young program to provide the "semantic clues" needed to identify the concepts contained in the text.

CHAPTER III. THE PRESENT RESEARCH

> Whenever you are asked if you can do a job, tell 'em "Certainly I can!"--and get busy and find out how to do it.
>
> Theodore Roosevelt

1. Phase 1: An Algorithm for Graphic Notation of English Sentences (AGNES)

> At this moment the King, who had been for some time busily writing in his note-book, called out "Silence!" and read out from his book "Rule Forty-two. All persons more than a mile high to leave the court."
>
> Everybody looked at Alice.
>
> "I'm not a mile high," said Alice.
>
> "You are," said the King.
>
> "Nearly two miles high," added the Queen.
>
> "Well, I sha'n't go, at any rate." said Alice: "besides, that's not a regular rule: you invented it just now."
>
> "It's the oldest rule in the book," said the King.
>
> Lewis Carroll, Alice's Adventures in Wonderland

1.1. The Basic Algorithm

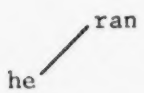
The initial phase of this research involved the definition of an algorithm to generate a graphic notation of English text. Each sentence in English is represented by a network comprised of nodes and edges. A sentence is viewed as being composed of one or more clauses, each consisting of a predicate and, optionally, a subject, one or more objects and/or modifiers. Each of these clause components is, when present in a sentence, represented in the network as a node. Relationships between nodes are represented as edges of the network. An edge may be labeled with a relational word. Several types of edge are employed;

they are defined as follows.

Def. 1. Subject-Predicate Edge:

The subject-predicate edge is a diagonal line from the subject to the predicate and has a slope of +1.

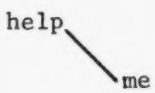
Example 1.



Def. 2. Predicate-Object Edge:

The predicate-object edge is a diagonal line from the predicate to the object and has a slope of -1.

Example 2.



Def. 3. Modifier Edge:

The modifier edge is a vertical line which depends the modifier from the word modified.

Example 3.

Def. 4. Conjunctive Edge:

The conjunctive edge is any dotted line.

Example 4.

Sally I

The definition of these edges constitutes a set of rules regarding the positioning of subject, predicate and modifiers within the network. The following rules complete the algorithm.

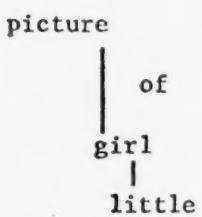
Rule 1. Articles:

Articles are not considered modifiers and are not displayed.

Rule 2. Prepositional Phrases:

Prepositional phrases are comprised of a labeled modifier edge, the modifier (object of the preposition) and possibly secondary modifiers.

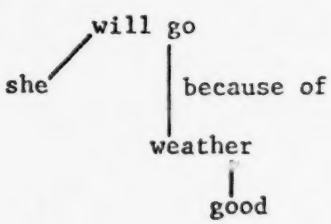
Example 5.



Rule 3. Compounds:

The auxiliary and main verbs constitute a single node. A compound preposition is considered a single edge-label.

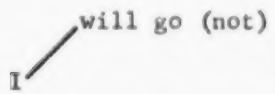
Example 6. She will go because of good weather.



Rule 4. Negations:

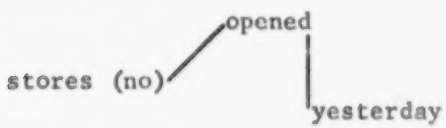
Negations are considered part of the verb or noun node rather than modifiers. The negating word is set off with parentheses.

Example 7. . I will not go.



Example 8. . I have no money.

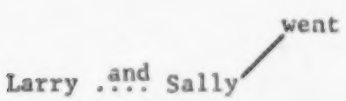
Example 9. No stores opened yesterday.



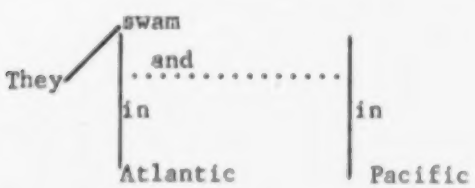
Rule 5. Coordinate Conjugation:

Compounds are connected horizontally and the conjunction labels the edge.

Example 10. Larry and Sally went.



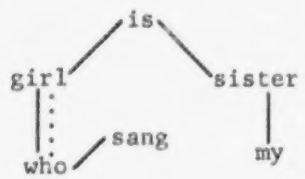
Example 11. They swam in the Atlantic and in the Pacific.



Rule 6. Relative Clauses:

If a dependent clause is a relative clause which renames a node of another clause, a modifier edge is drawn terminating at the relative pronoun node. (The dotted edge in example 12 is explained in Rule 9.)

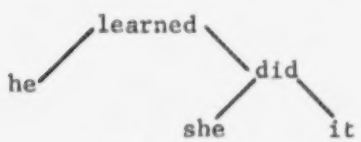
Example 12. The girl who sang is my sister.



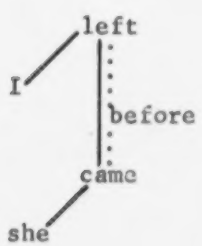
Rule 7. Dependent Clause (Normal Word Order):

If a dependent clause is in normal word order (noun-predicate-noun) and Rule 6 does not apply, the appropriate function edge terminates at the predicate node of the dependent clause.

Example 13. He learned she did it.



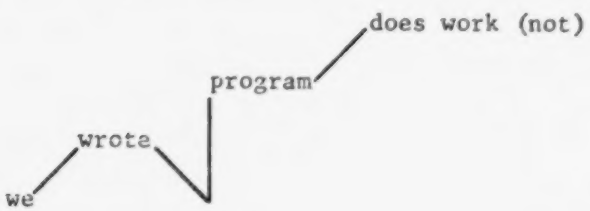
Example 14. I left before she came.



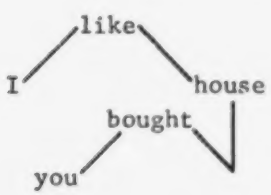
Rule 8. Dependent Clause (Reverse Word Order):

If a dependent clause is signaled by a reverse word order (noun-noun-predicate) the first noun is assumed to have position within the parent clause. The edge is drawn to a null node in the dependent clause which is attached to the dependent predicate by a predicate--object edge.

Example 15. The program we wrote does not work.

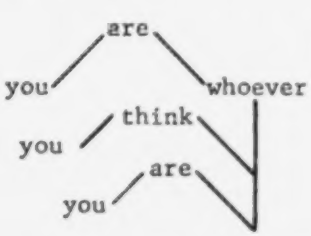


Example 16. I like the house you bought.



This rule may be extended to several levels when necessary. Thus,

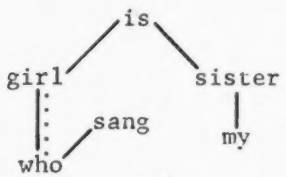
Example 17. You are whoever you think you are.



Rule 9. Subordinate Conjugation:

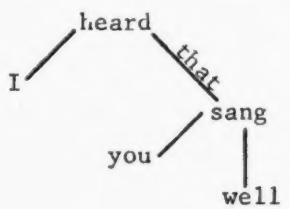
If the dependent clause is joined to the main clause by a conjunction or pronoun, the function edge is a double edge, composed of the clause function edge and a conjunctive edge.

Example 18. The girl who sang is my sister.

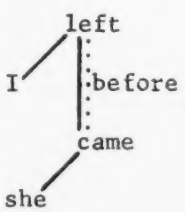


If the pronoun or conjunction does not occupy a node in the dependent clause, it labels the double edge.

Example 19. I heard that you sang well.



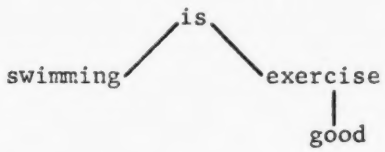
Example 20. I left before she came.



Rule 10. Single-Word Verbals:

Single-word verbals are not considered clauses and therefore merit no special treatment.

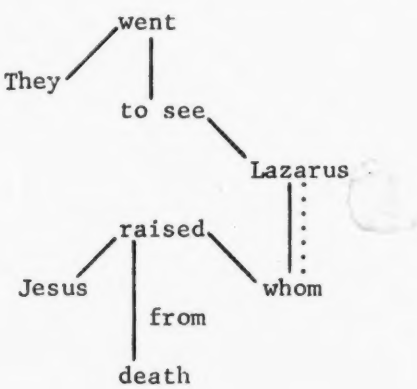
Example 21. Swimming is good exercise.



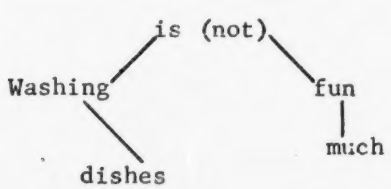
Rule 11. Verbal Phrases:

Verbals which have a subject or an object are considered to be clauses.

Example 22. They went to see Lazarus whom Jesus raised from death.



Example 23. Washing dishes is not much fun.



With these rules, a graph may be generated for any English sentence. The graphs of sentences from a short article are included in Figure 3.1 to

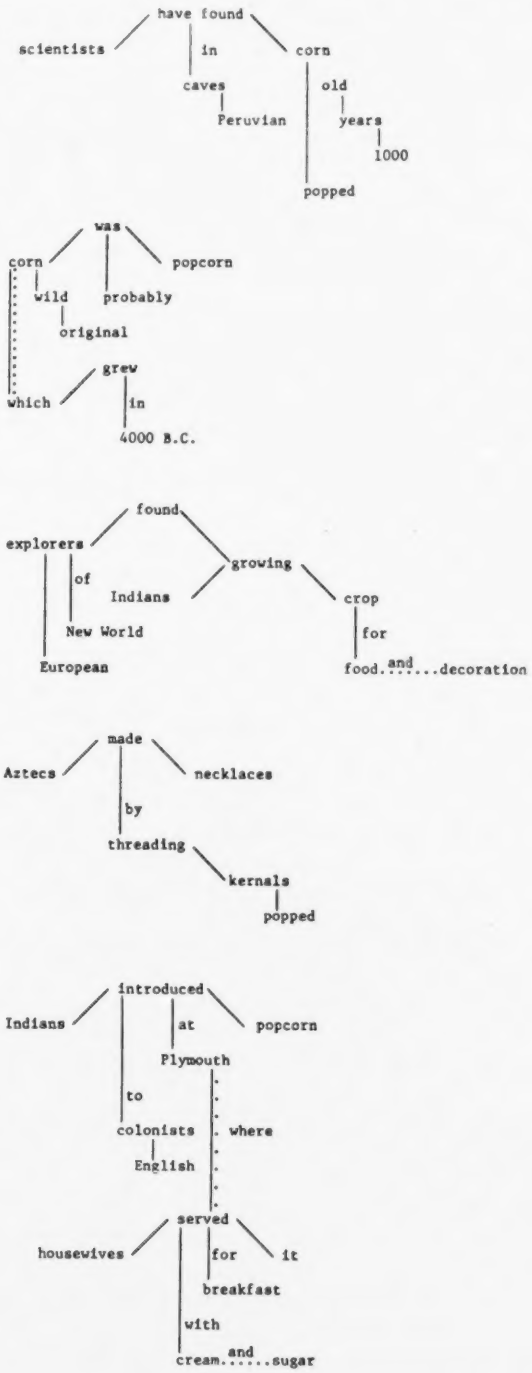


Figure 3.1 AGNES graphs of "Age-Old Popcorn" (d1).

further illustrate these rules.

For certain applications, it may be desirable to specify relator types, to indicate implied relationships, or to reduce the amount of data displayed. These extensions to the basic algorithm are defined in the section which follows.

1.2. Display Extensions

1.2.1. Specification of Relator Types

Within the case grammar framework[1] verbs may be classified as follows: stative, action, process or action-process. Intuitively, non-stative verbs are those which can be used in a sentence to answer the question "What happened?" Stative verbs are those which cannot be used to answer that question. Action verbs answer the question "What did N do?" where N is a noun in the clause. Process verbs answer the question "What happened to N?" where N is a noun in the clause. Action-process verbs answer both the preceding questions. For example, the question "What happened?" cannot be answered by the statement "The balloon is red." In this sentence "is" is a stative verb. "He ran" describes an action and answers the question "What did he do?" Thus "ran" is an action verb. "The balloon broke" answers the question "What happened to the balloon?" so in that statement, "broke" is a process verb. If the sentence had been "He broke the balloon" it would answer both "What did he do?" and "What happened to the balloon?" thus "broke" in this case is an action-process verb.

---

1. This discussion owes a great deal to Chafe (26).

A stative verb requires an object. An action verb requires an agent. A process verb requires an object and an action-process verb requires both an agent and an object. These combinations constitute the four case frames shown in Figure 3.2. Cook (31) has combined each of these with experiencer, beneficiary or locative nouns to produce the 16-case frame matrix depicted in Figure 3.3.

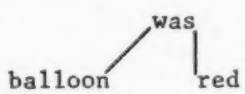
The Young analysis defines locative as a peripheral case (signaled by a preposition). Therefore it will not be considered in this set of rules which are concerned only with major cases.

The major case frames are specified by directional edges in the AGNES graphs as follows:

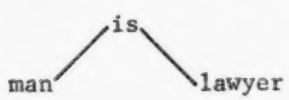
Rule 12. Stative Verb:

A stative verb [0] or [0,0] dictates an undirected edge to the object.

Example 24. The balloon was red.



Example 25. The man is a lawyer.



Rule 13. Action Verbs:

The edge linking an action verb [A] to the agent is directed from the agent to the predicate.

BASIC VERB TYPE

1. State [O]

The cup is white

2. Process [O]

The cup broke.

3. Action [A]

The boy sings.

4. Action-Process [A, O]

The boy closes the book.

Figure 3.2 The basic verb types described by Chafe (26).

BASIC VERB TYPES	EXPERIENCER	BENEFACTIVE	LOCATIVE
1. State [O] is (broken) was (dry)	1. State Experiencer [E,O] know like	1. State Benefactive [B,O] have own	1. State Locative [O,L] is (in) was (on)
2. Process [O] break dry	2. Process Experiencer [E,O] feel hear	2. Process Benefactive [B,O] find lose	2. Process Locative [O,L] come go
3. Action [A] dance laugh	3. Action Experiencer [A,E] please answer	3. Action Benefactive [A,B] bribe help	3. Action Locative [A,L] run walk
4. Action-Process [A,O] kill break	4. Action-Process Experiencer [A,E,O] ask tell	4. Action-Process Benefactive [A,B,O] buy give	4. Action-Process Locative [A,O,L] put take

Figure 3.3 The case frame matrix developed by Cook (31).

Example 26. He ran.

Rule 14: Process Verbs:

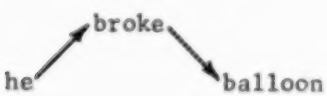
The edge of a process verb [O] is directed from the predicate to the object. If no beneficiary, agent or experiencer cases are present, the edge is bi-directional, that is, from the object to the verb and from the verb to the object.

Example 27. The balloon broke.

Rule 15: Action-Process Verbs:

The edge of an action-process verb [A,O] is directed from the agent to the predicate and from the predicate to the object.

Example 28. He broke the balloon.



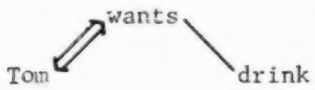
Rule 16: Experiencer Frames:

An experiencer frame is depicted as a double-edged arrow (⟸⟹) directed from the predicate to the experiencer and from the experiencer to the predicate.

Example 29. I thirst.

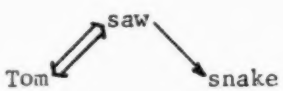
In a stative [E,O] frame, only the experiencer edge is directional.

Example 30. Tom wants a drink.



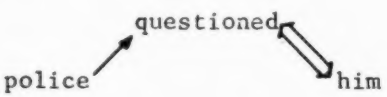
In a process [E,O] frame, the predicate-object edge is directed from the predicate to the object.

Example 31. Tom saw a snake.



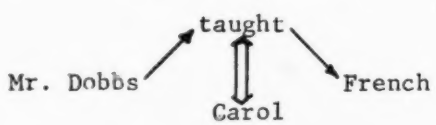
In an action [A,E] frame, the subject-predicate edge is directed from the agent to the predicate.

Example 32. The police questioned him.



In an action-process [A-E-O] frame the subject-predicate edge is directed from the agent to the predicate and the predicate-object edge is directed from the predicate to the object.

Example 33. Mr. Dobbs taught Carol French.

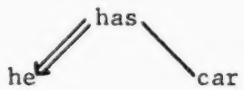


Rule 17. Beneficiary Frames:

> Beneficiary frames are indicated by a double-edged arrow directed from the predicate to the beneficiary.

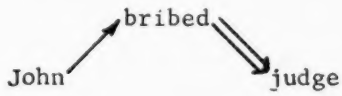
In the stative [B,O] frame, the predicate-object edge is non-directional.

Example 34. He has a car.



In the action [A,B] frame, the subject-predicate edge is directed from the agent to the predicate.

Example 35. John bribed the judge.

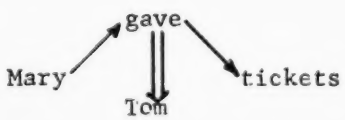


In the process [B,O] frame, the predicate-object edge is directed from the predicate to the object.

Example 36. He lost the tickets.

In the action-process [A,B,O] frame, the subject-predicate edge is directed from the agent to the predicate and the predicate-object edge is directed from the predicate to the object.

Example 37. Mary gave Tom the tickets.

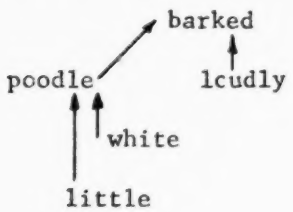


These modifications of the verb-related edges help to specify the major relationships within the clause; that is, noun-verb relationships. Rules for alternation of the secondary relationships, the modifier edges, follow:

Rule 18. Single Word Modifiers:

The edge of a single word modifier (adjectives or adverbs) is directed from the modifier to the word modified.

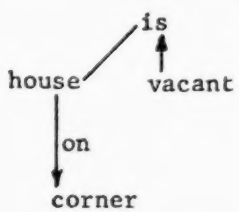
Example 38. The little white poodle barked loudly.



Rule 19: Modifier Phrases:

The edge of a modifier phrase is directed from the node modified to the modifier.

Example 39. The house on the corner is vacant.



These extensions to the basic algorithm differentiate relator types within a functional relationship[2] on the basis of case roles and syntactic form. The next section describes a set of rules whereby some of the non-functional relationships implied by the syntax within a sentence may be identified.

---

2. A functional relationship is defined to be the relationship specified by an edge between two nodes.

1.2.2. Identification of Implied Relationships

There are probably many relationships among elements of a sentence which are implied by the syntax, ordering or case roles of the elements of a sentence, which are not direct functional relationships. The following set of rules identify two such classes of relationships and describe the secondary relator edges which are used to explicate them in the AGNES graphs.

1.2.2.1. Stative Verbs

Among the stative verbs in the stative [0] and stative [0,0] frames, several types may be distinguished.

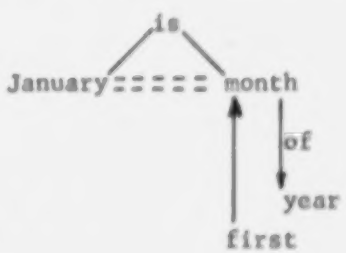
Rule 20. Equivalence:

> An equivalence relationship is recognized by a definite article or proper noun in both the "subject" and "predicate nominative" (both objects).

Example 40s. January is the first month of the year.

The equivalence relationship is depicted as a double dashed line.

Example 40g.



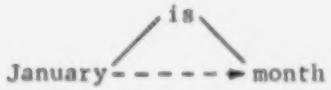
Rule 21. Classification:

> A generic-specific relationship is recognized by a definite article or proper noun in one object slot and an indefinite article in the other.

Example 41a. January is a month.

In this example, "January" is a member of the set of all months. This classification is depicted as a dashed line directed from the definite to the indefinite node. A ---➤B is equivalent to the set notation A ∈ B.

Example 41g.



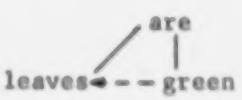
Rule 22. Modification:

> Modification is peculiar to the stative[O] frame and involves what are commonly called "predicate adjectives."

Example 42a. The leaves are green.

Modification is depicted as a single dashed edge directed from the modifier to the object.

Example 42g

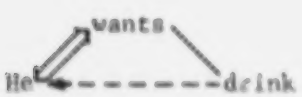


The stative [E,O] and [B,O] frames suggest direct relationship of E-O and B-O.

Rule 23. Experiencer/Stative Frame:

> In the experiencer/stative frame, a dashed edge directed from the object to the experiencer is used to demonstrate the experience relationship.

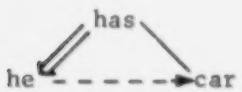
Example 43.



Rule 24. Beneficiary/Stative Frame:

In the beneficiary/stative frame, a dashed edge directed from the beneficiary to the object shows the possessive relationship.

Example 44.



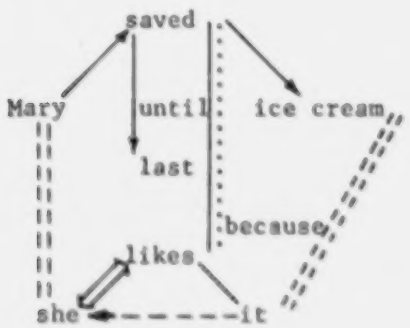
1.2.2.2. Pronouns

Another group of implied relationships are those signaled by a pronoun.

Rule 25. Personal Pronouns:

A personal pronoun and its antecedent constitute an equivalence class. A personal pronoun and its antecedent are connected by a double dashed edge.

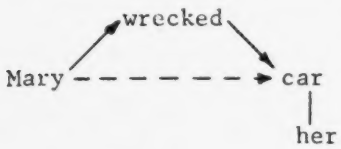
Example 45 Mary saved the ice cream until last because she likes it.



Rule 26. Possessive Pronouns:

The possessive form of a pronoun is connected to its antecedent by a single dashed edge directed from the antecedent to the "thing possessed."

Example 46. Mary wrecked her car



The preceding set of rules governing implied stative and pronominal relationships constitute a sample of the type of display which might be developed to exhibit implied syntactic relationships.

1.2.3. Reductions

A third extension to the basic algorithm is a reduction procedure. Words which are linked by an equivalence relation or which otherwise seem unnecessary to the "meaning" of a sentence may be eliminated from the graph. Only a few rules are suggested, however, because it is usually not desirable to conceal data, and in any case, it has been decided in the development of this research that the initial structure developed by AGNES should be maintained. Many of the rules presented thus reflect this decision.

Rule 27. Possessive Adjectives:

> Possessive pronominal adjectives are eliminated if the antecedent is in the same clause, and if a possessive edge (---➤) has been drawn from the antecedent to the thing possessed (by Rule 26).

Example 47.

Rule 28. Personal Pronouns:

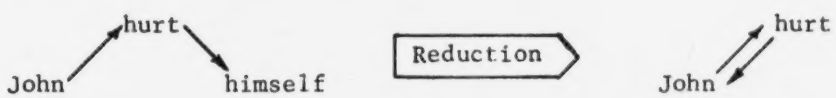
Personal pronouns which are joined by an equivalence edge (==) to their antecedent (by Rule 25) are erased. The node is not eliminated.

Example 48. John is the captain because he is tall.

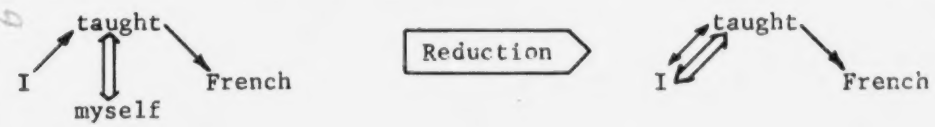
Rule 29. Reflexive Pronouns:

Reflexive pronouns are eliminated, but the mirror image of the edge relating the pronoun to the predicate must be drawn from the antecedent to the predicate.

Example 49. John hurt himself.



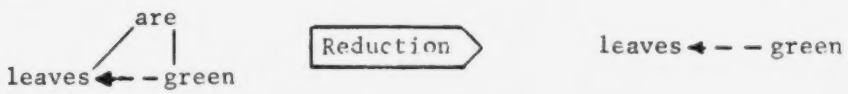
Example 50. I taught myself French.



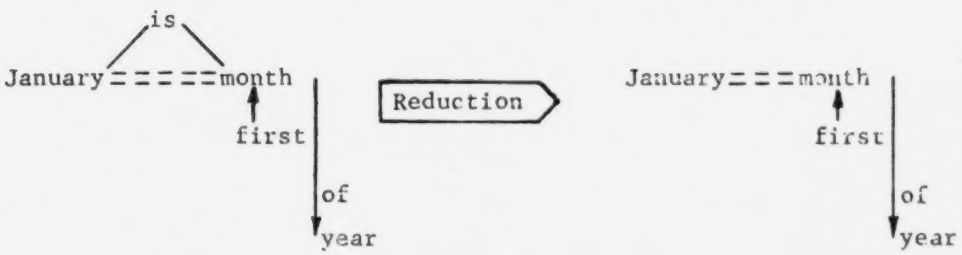
Rule 30. Stative Verbs:

Verbs of the stative [0] and stative [0,0] case frames are eliminated.

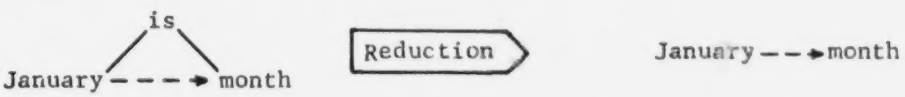
Example 51. The leaves are green.



Example 52. January is the first month of the year.



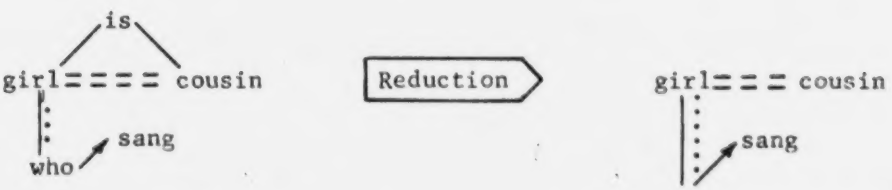
Example 53. January is a month.



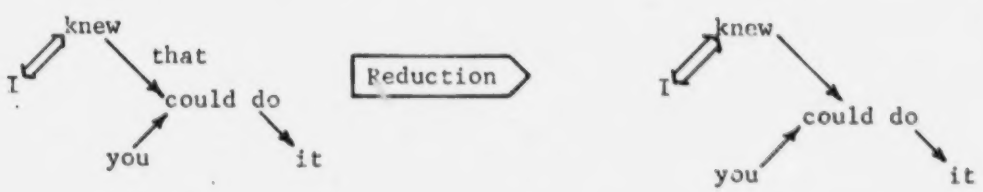
Rule 31. Relative Pronouns:

Relative pronouns and the word "that" are erased. The nodes are not eliminated.

Example 54. The girl who sang is my cousin.



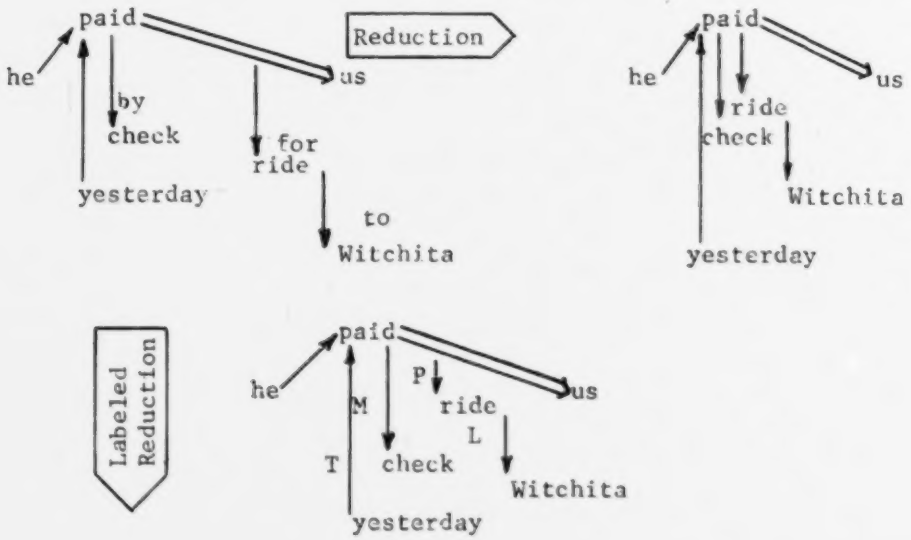
Example 55. I knew that you could do it.



Rule 32. Labels:

> Labels are erased or replaced where applicable by peripheral case markers. In that case, the locative case is denoted by "L"; time by "T"; manner by "M"; accompaniment by "A"; purpose by "P" and cause by "C".

Example 56. He paid us yesterday by check for the ride to Witchita.



A reduced graph may be particularly helpful in comparing one segment of text with another.

A method of generating a graph for an English sentence has been described above together with several options for extended display capabilities. Figure 3.4 illustrates application of the extended rules to the graphs of Figure 3.1. If one supposes that an entire document might yield a network, then a method of connecting sentence graphs to form such a network is needed.

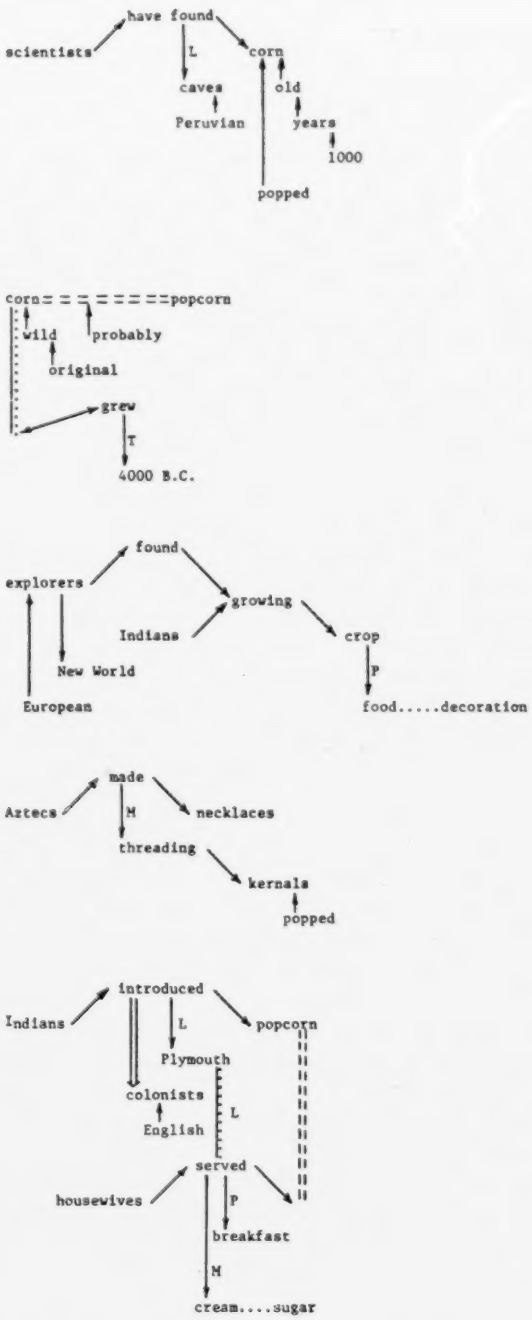


Figure 3.4 Application of the extended AGNES algorithm to the text of Figure 3.1

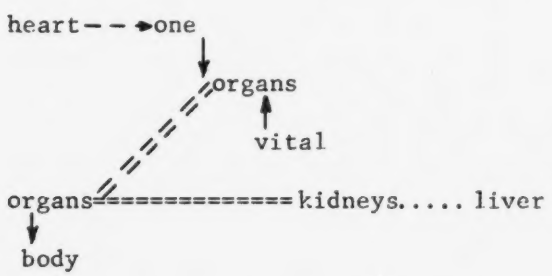
1.3. Intersentence Relationships

In order to conceptualize the type of network which might be built, suppose that each clause forms a plane. Now imagine that these planes intersect at common nodes and that the planes are connected by specified or implied relations. A three-dimensional network of this sort is difficult to display, but an attempt is made in Figure 3.5. A set of rules for two-dimensional display of this model has been developed. The rules are as follows.

Rule 33. Repetitions:

Two or more occurrences of the same word (assuming they occupy nodes) are linked in sequence with a double dashed edge.

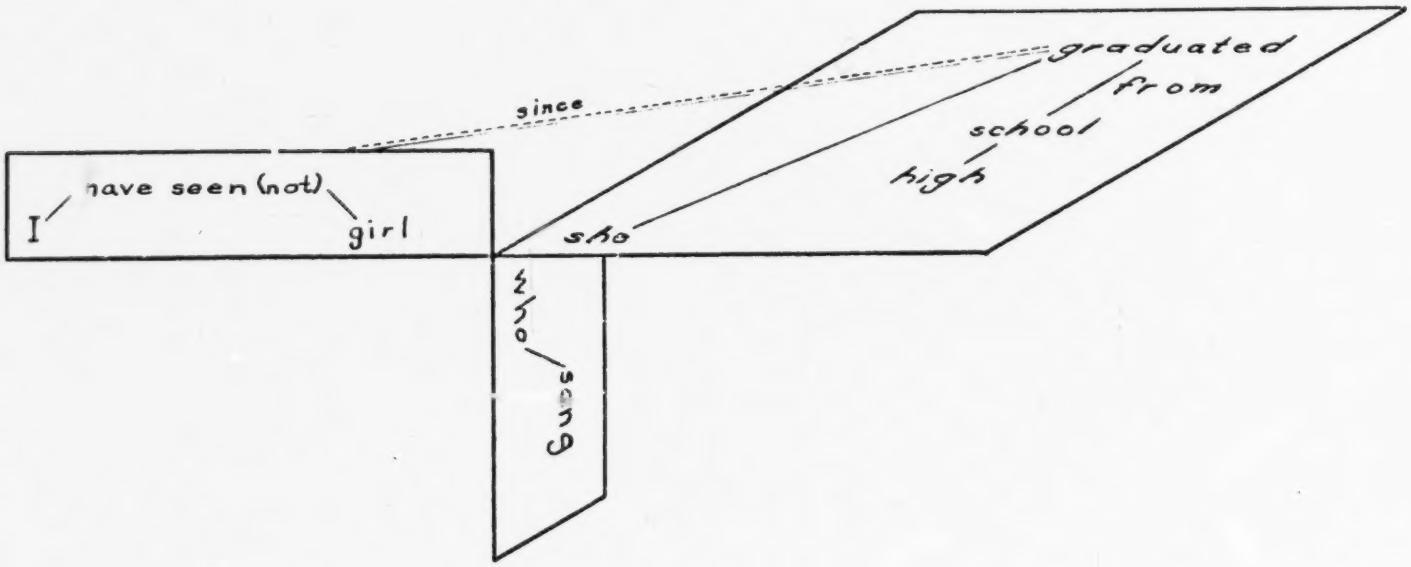
Example 57. The heart is one of the vital organs. Other organs of the body are the kidneys and liver.



The graphs used in these examples are in reduced form.

Rule 34. Antecedents:

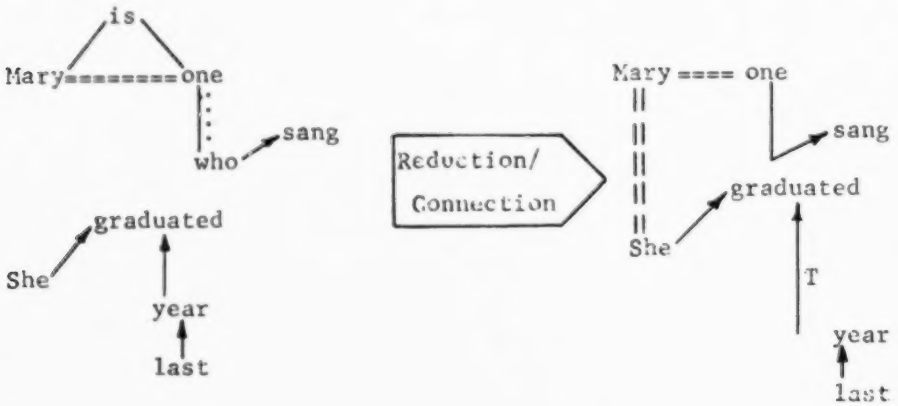
Antecedents of personal pronouns will be linked to the pronoun with a double dashed edge.



I have not seen the girl who sang since she graduated from high school.

Figure 3.5 Representation of the three-dimensional surrogate-model network.

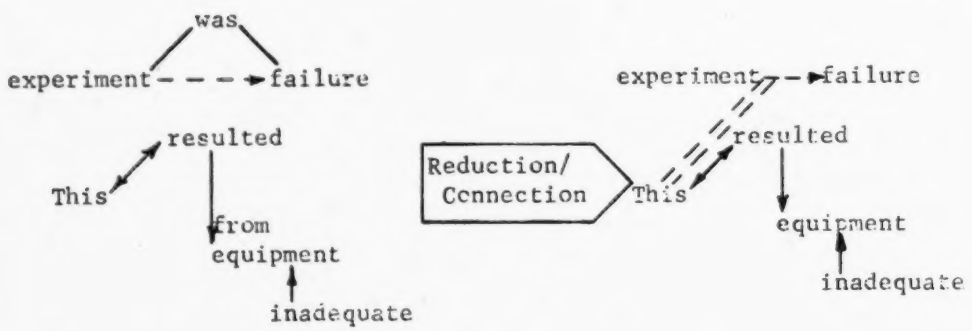
Example 58. Mary is the one who sang. She graduated last year.



Rule 35. Demonstratives:

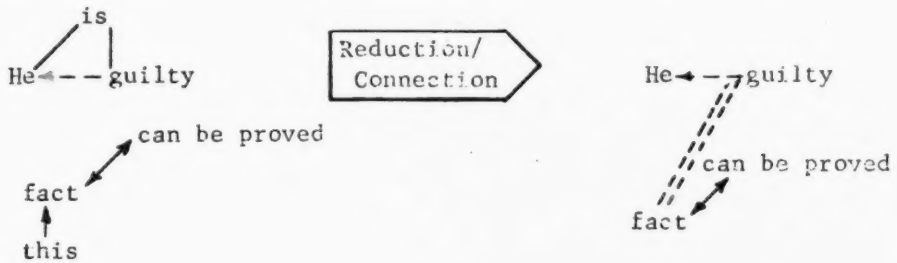
A demonstrative pronoun is linked to the predicate of the preceding clause with a double dashed edge.

Example 59. The experiment was a failure. This resulted from inadequate equipment.



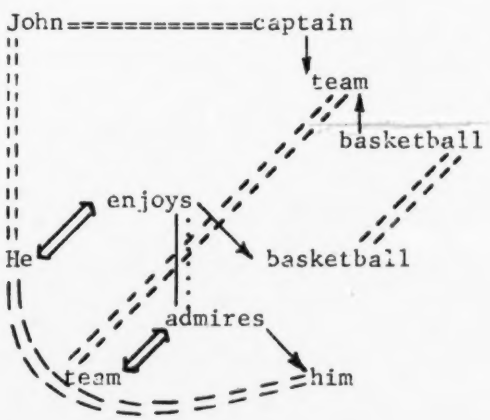
A demonstrative adjective is linked to the predicate of the preceding clause if the node it modifies is not linked by Rule 33 or Rule 34 to another sentence.

Example 60. He is guilty. This fact can be proved.



In some cases, intersentence-relator edges will curve around or overwrite nodes. Either technique is acceptable.

Example 61. John is captain of the basketball team. He enjoys basketball, because the team admires him.



It would be naive to assume that there are no problems with automatic identification of pronoun antecedents. The traditional grammars suggest that "a pronoun generally agrees with its antecedent in gender, person and number (32)". And usually, the antecedent may be defined as the most recent noun which agrees with the pronoun in gender, person and number. Unfortunately, the following problems are inherent in the attempt to automate such a definition:

- Gender (at least in English) cannot be ascertained automatically.
- Compounds and certain connectives may change the number.

  Paul left with Jack. They were upset...

- The word "it" may refer to a positive clausal or a neuter antecedent.

  They asked him to speak. Did he do it?

  The dog was cute. It sat up and begged.

- Direct quotation changes the person.

  He told them, "You may come."

- Possible antecedents may be eliminated due to logical impossibility.

  Harmon took aim and shot Kunz through the heart.

  Then he moved along the east side of the church...

- Perhaps the more difficult problem with which to contend is the use of other nouns to establish intersentence relationships. The following example admirably illustrates the problem.

> The unprofessional fern collector is likely to agree with Gray in considering the Adder's Tongue "not common." Many botanists, however, believe the plant to be "over-looked rather than rare." In an article on O. vulgatum which appeared some years ago in the Fern Bulletin, Mr. A. A. Eaton writes: "Previous to 1895 Ophioglossum vulgatum was unknown to me and was considered rare. Early in the year, a friend gave me two specimens. From these I got an idea of how the thing looked. On the 11th of last July, while collecting H. lacera

> in a bound-out mowing field, I was delighted to notice a spike of fruit in the grass. A search revealed about sixty; just right to collect, with many unfruitful specimens. A few days later , while raking in a similar locality, I found several, within a stone's throw of the house, demonstrating again a well-known fact that a thing once seen is easily discovered again. On the 23rd of August, while riding on my bicycle, I noticed a field that appeared to be the right locality and an investigation showed an abundance of them. I subsequently found it in another place. (33)

In addition to the pronominal antecedent difficulties there may be some weaknesses in the clausal relationship system. For example, it might be desirable to provide linkages between adjectival and nominal forms of the same word. This is not possible with the system as it is presently defined.

The present research does not ignore these difficulties. Rather the aim is to do as well as possible without attempting to overcome them. Figure 3.6 displays the graphs of Figure 3.4 connected by inter-sentence relators to form a network of the text.

2. Phase 2: Experimentation and Testing

> One day the mice held a general council to consider what they might do to protect themselves against their common enemy, the Cat. Some said one thing and some said another, but at last a Young Mouse stood up and announced that he had a plan which he thought would solve the problem.
>
> "You will all agree," said he, "that our chief danger lies in the unexpected and sly manner in which our enemy comes upon us. Now, if we could receive some warning of her approach, we could easily hide from her. I propose, therefore, that a small bell be obtained and attached by a ribbon to the neck of the Cat. In this way we could always know when she was coming and be able to make our escape."

Figure 3.6 Application of the intersentence relator rules to the graphs of Figure 3.4.

> This proposal was met with great applause, until an Old Mouse arose and said, "This is all very fine, but who among us is so brave? Who will bell the Cat?" The mice looked at one another in silence and nobody volunteered.
>
> It is easier to suggest a plan than to carry it out.
>
> Aesop's Fables, "Belling the Cat"

The second phase of the present research is the application of the above procedures to text samples. Several types of documents were selected, including an abstract, a short technical paper and non-technical articles of several subjects. (Several examples from the set are included as Appendix A to this thesis). The networks for each of the articles were drawn manually in accordance with the basic algorithm (Section 1.1.).

On studying the networks, several trends can be discerned. First, the graphs tend to "cluster", that is, there exist multiple references to the same word throughout a text or portion of the text. Second, case roles appear to be consistent within a portion of text and third, case roles seem to parallel "intellectual" analysis. For example, in Figure 3.7 "church" is seen to be a highly connected node, which has been termed a "cluster". One might suppose that the article is "about" churches. However, the case role of the word "church" is clearly locative in the majority of instances in this text. An "intellectual" analysis reveals that indeed "church" is only descriptive of the location of the incident being reported. That the case role is consistent over a portion of text and that "intellectual" analysis confirms the consistency of the case assignment are two developments which may prove a valuable aid in automatic determination of the topic

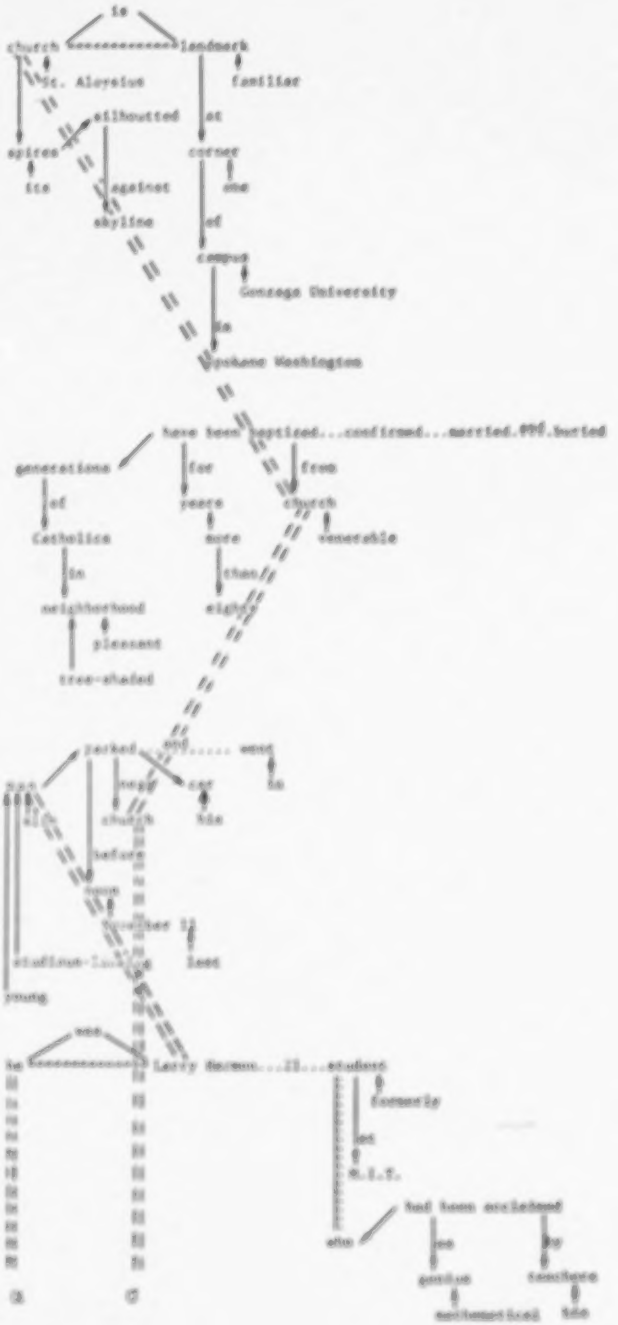


Figure 3.7 ACNES graph of a portion of "Larry's Trip to Tragedy" (d2).

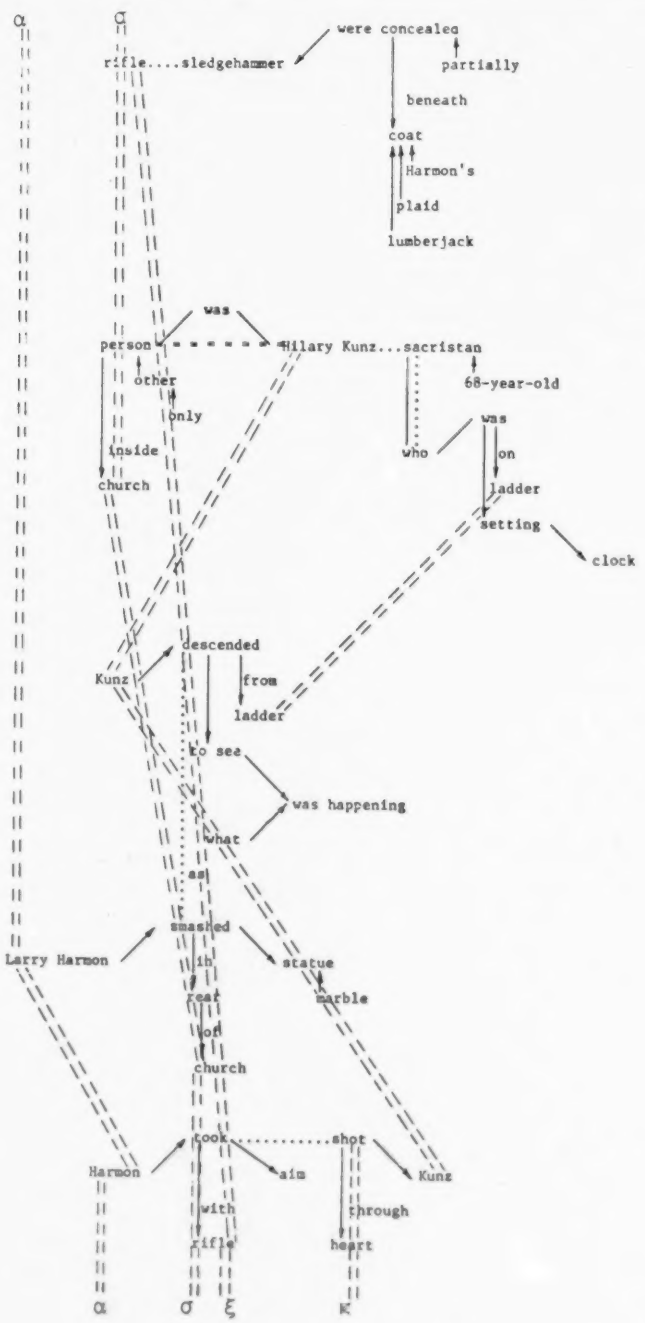


Figure 3.7 (continued).

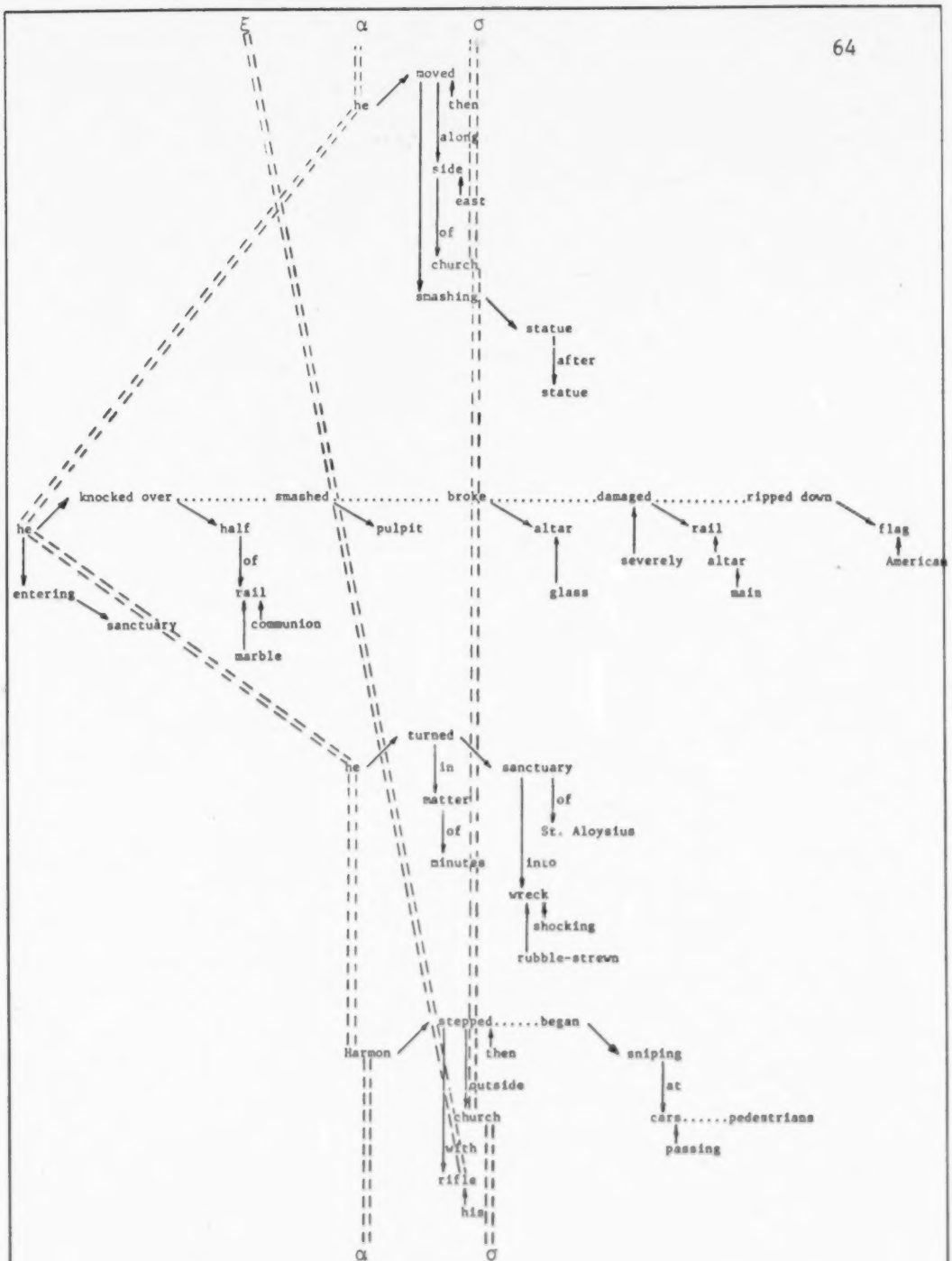


Figure 3.7 (continued).

Figure 3.7 (continued).

Figure 3.7 (concluded).

of a text.

Further evidence of these developments can again be found in Figure 3.7. There is a cluster at the node "Larry." The word "Larry" usually occurs as an agent. "Intellectually", Larry is clearly the "doer" in the text. The relationships of "Larry" and "church" may thus be defined in terms of case assignments.

Not only do case roles help to relate terms within a text, they may help to distinguish between texts. Notice the graph of the "Age-Old Popcorn" article (Figure 3.6). The major cluster is "popcorn" and it is not not strongly connected to any other nodes. "Popcorn" is in the object case, but since there is no clear agent with which to relate it, one might be tempted to ignore this fact. To do so would be to lose a valuable distinction between this article and one, say, about the physical properties or effects of popcorn (for example, "popcorn pops at..." or "popcorn causes...").

A more detailed description of these results is found in Chapter IV of this thesis. The third phase of the research is to demonstrate the feasibility of computer generation of the proposed system.

3. Phase 3: Computer Feasibility: Design Considerations

> He then led me to the frame, about the sides whereof all his pupils stood in ranks. It was twenty foot square, placed in the middle of the room. The superficies was composed of several bits of wood, about the bigness of a die, but some larger than others. They were all linked together by slender wires. These bits of wood were covered on every square with papers pasted on them and on these papers were written all the words of their language in their several moods, tenses and declinsions, but without

> any order ... The pupils at his command took each of them hold of an iron handle, whereof there were forty fixed round the edges of the frame, and giving them a sudden turn, the whole disposition of the words was entirely changed.
>
> Jonathan Swift, Gulliver's Travels

In order to demonstrate computer feasibility, it is necessary to describe a design for such a program. The program itself is beyond the scope of this research, but the logic has been developed, and reasonably detailed flowcharts are included in Appendix B.

The programs developed by Young operate on an IBM 370/165 but the 370 machine configuration at The Ohio State University has no graphic display capability. Thus the following discussion will assume the use of an IBM 1130 with a 2250 display unit.

Three steps must be included in the design of a program to implement AGNES. First, given the grammatical class of each word, and the identity and type of each phrase and clause in the sentence, assign relator edges as specified in the algorithm with minimum storage allocations. Second, order the construction of the graph according to the results of the first step. And third, draw and label the graph.

3.1. Assignment and Storage of Edges

The first step is external to graphic considerations and could be appended to the Young programs operating on the 370. The output of this program would probably be the mode of storage for the document surrogates.

The assignment of relator edges must completely specify the graph. An efficient approach is demonstrated in Table 3.1. Associated with each word in the sentence are two data items; the edge and the word it

Table 3.1 Storage of Edge Assignments for Graphic Surrogates

Word Order	Word in Sentence	Edge	Edge Code*	Reference Word	Reference Code**
1	Scientists	/	1	have	2
2	have			found	3
3	found	\	2	corn	5
4	popped	|	3	corn	5
5	corn				
6	1000	|	3	years	7
7	years	|	3	old	8
8	old	|	3	corn	5
9	in	|	3	found	3
10	Peruvian	|	3	caves	11
11	caves	†	5	in	9

*Edge Codes: 1 / , 2 \ , 3 | , 4 ... , 5 †
**Refers to word order

links. These data may be represented by numeric codes to save storage requirements. Since, in the resultant graph, a single word may be associated with more than one edge, a decision must be made as to which edge is assigned to that word. Assignments are made according to the following rules.

- A subject signals the subject-predicate edge (/) and refers to the predicate.
- A predicate signals the predicate-object (if applicable) edge (\) and refers to the object.
- A preposition signals the terminal modifier edge (|) and refers to the modified word.
- An object of the preposition signals a non-terminal edge (↑) and refers to the preposition.
- Other modifiers signal the terminal modifier edge (|) and refer to the word modified.
- Subordinate conjunctions signal the double connective edge (·····) and refer to the dependent verb.
- Correlative conjunctions signal the single conjunctive edge (···) and refers to the first word coordinated.
- Conjunctive phrases constitute a new entry in the table consisting of the conjunction, the conjunctive edge and the second word coordinated.
- Subordinate clause markers constitute a new entry in the table consisting of the word in the dependent clause which is to be connected (according to AGNES rules), the edge as specified above,

and the word in the main clause to which the clause is appended.

- An auxiliary verb receives no edge assignment and refers to the next auxiliary or to the main verb. The subject references the first auxiliary, the main verb the object.
- The first word of a compound preposition is assigned the terminal modifier edge (|) and refers to the word modified. The second and subsequent words are assigned no edge but each refers to the word immediately preceding. The object of the preposition references the last preposition.

Words which do not fit the above categories are assigned no edge nor reference.

The resulting table (Table 3.1) completely specifies the graph at small expense in storage.

It is not possible, however, to construct the graph from the table in a sequential manner. For example in the sentence of Table 3.1 "Scientists have found popped corn 1000 years old in Peruvian caves," "1000" is encountered with a reference to "old" before "old" exists on the graph. There is no place from which to depend "1000". This problem is the impetus for the second step of the program design, ordering the construction.

3.2. Ordering the Construction

The following ordering, schematized in Table 3.2, is suggested.

Beginning with the first element of the independent clause, draw the word-edge-word triple.

Table 3.2 Ordering Construction of a Graph from the Storage Table*

Word in Sentence	Edge	Reference Word	Clause Reference	Construction Order
The				
original	|	wild		1
wild	|	corn		2
corn	/	was		3
which	/	grew		8
grew				12
in	|	grew		9
4000	+	in		10
B.C.		4000		11
was	\	popcorn		5
probably	|	was		4
popcorn				6
which	|:		corn	7

* Beginning with the independent clause, the triples of all occurrences of a word in the reference column (code word) are processed first. Then the triple of the code word in the sentence column is drawn, and its reference becomes the new code word.

A. Mark the sentence position (left column) of the reference word (center column).

B. Search the reference column for another occurrence of the reference word. If it appears in the reference (center) column, mark it and draw the new word-edge pair.

C. Now using the sentence position word (left column) of this triple search the reference column. If an occurrence of the sentence position word is found in the reference column, draw that pair and repeat from C (search for new sentence word). Continue until there is no reference for the sentence word.

Return to the reference column marker and search the reference column for the next occurrence of the marked word. If one is found, repeat from B.

When the entire reference column has been searched, return to the marker in the sentence position column. Draw the edge-word pair and repeat from A.

The clause is complete when a blank edge and reference appear for the marked sentence position.

If the sentence contains a dependent clause, draw the pair specified by the clause triple. Search the sentence position column for the regular occurrence of the clause marker. Draw the pair specified and return to A.

A flowchart of the above procedures is included in Appendix B. This ordering enables the graph of any sentence to be drawn from the table

described in Step One.

3.3. Graphic Considerations

The last step in the design plan is the actual graphic considerations. These may be summarized briefly as

a) What do we draw?

b) Where do we draw it? and

c) What screen information is needed to repeat a and b?

3.3.1. What to Draw

Character data on the 2250 is a fixed size and spacing. A maximum of 52 lines of 74 characters each may be displayed at one time, where character spacing is 14 raster units and line spacing is 20.

Assuming that we wish to "double space" the graphs, vertical edges must be 40 raster units and diagonal lines the hypotenuse of a 40 X 40 right triangle. Since we must have the ability to draw the edge from either direction, and since 8 edges are possible (4 solid and 4 dotted) a total of 16 vectors are required. The graphic routines to produce these vectors may be programmed with incremental instructions so that the same routine may be used to draw the edge anywhere on the screen.

3.3.2. Where to Draw It

Where to position the line or character string is the next graphic consideration. The following guidelines are suggested.

- A diagonal line beings at the base of the character space after the last character of the first word. The second character string beings with the next character space after the edge.

- A terminal vertical edge beings at the base of the last character of the word modified. If an edge is already appended there, the new edge is positioned one character to the left until an empty "slot" is found. The modifier begins one character space to the right of the edge. If a word appears within the character span of the modifier, the edge is extended another line until the modifier "fits."
- Non-terminal modifier edges are drawn as continuations of the modifier edge. The object of the preposition follows the "modifier" rules of above.

3.3.3. What to Retrieve

What screen information is necessary to carry out these guidelines is the last graphic consideration to be discussed. It is necessary to "remember" where a word has been written on the screen so that

a) other words may be linked to it and

b) words will not be written over one another.

Sufficient information to be retrieved from the screen are the X and Y coordinates of the base point of the character space before and after each word. The graph of Figure 3.8 shows these points. Where to append modifiers may be calculated by subtracting one character space from the end point of the word modified. Existence of an edge in that "slot" may be determined by checking the table for an equal X coordinate. Once an empty "slot" has been located, the edge is extended line by line until the distance between the edge and any point on that line (to the right of the edge) is greater than the length of the modifier.

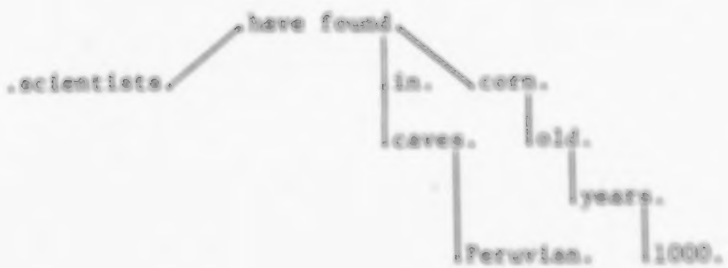


Figure 3.8 Diagram of graphic display showing points which must be returned from the display screen.

These considerations are the basics in graphic programming. These three steps outline the design of a set of programs which could be used to implement the basic AGNES algorithm, given the output from the Young programs. Computerization is feasible. Furthermore, the additional storage requirements are not extensive. The graphic program described is minimal and is meant to demonstrate feasibility rather than proficiency or sophistication.

As the final phase of the present research, consideration is given to the practical application of surrogate storage in natural language processing.

6. Phase 4: Investigation into Applications

> Without a further goal than syntactic analysis for its own sake, we are limited to judging these programs by some arbitrary nonoperational criteria of elegance, explanatory power and simplicity. A sufficient proof of the goodness of any of these theories lies in its usefulness for further processing.
>
> Daniel Bobrow

The worth of any of any system in its own right is, in my opinion, a moot question. It is only as the technique is applied that it can be judged. Accordingly, there follows a brief discussion of possible applications for the structural surrogates developed in this research.

6.1. Indexing

In the realm of indexing, several techniques are currently employed. It is proposed that each of these would benefit by using the structural surrogate as a document base.

Term indexes could be produced automatically using high connected nodes as potential "keywords". Notice that this is not a frequency approach. In the network of "Larry's Trip to Tragedy" (Figure 3.7) for example, the word "Larry" is used in only 6 of 11 references to the person. Other references are as pronouns or synonyms. In contrast, the network clearly links all references to Larry with an equivalence edge. Only major-case nodes would be selected for a term index; location, time, and other peripheral cases would be eliminated.

To produce a KWIC index, keywords could be taken in the context of the "most connected" sentence(s) of the text. In the "popcorn" graph (Figure 3.6) this might be the clause "Indians introduced popcorn to the English colonists at Plymouth" since both "popcorn" and "Indian" form clusters. It is suggested that a sentence selected in this manner would offer a good alternative to use of the title of a document for KWIC indexing.

Phrase indexes such as articulated indexes could also be automatically derived from the structural surrogate. Clusters are linked to other clusters by means of some relator (verb, preposition, etc.). These relationships are clearly specified in the surrogate and could be used to index the text. Note that with a broader definition of index phrase as a node-relator-node triple, verbs may be included in the phrase index entry.

Derivatives of present indexing techniques might also be explored. For example, consider the articulated index entry, "discovery of oil in Alaska". Chafe's rules for case assignment (26) can easily be applied

to the sentence from which this entry was derived, namely "Oil has been discovered in Alaska." It is discovered that an "action" of discovery is described; the "agent" is not known; the "object" is oil and the "location" is Alaska. Now, rather than permuting the phrases at articulation points such as

oil, in Alaska, discovery of

organize an entry in tabular fashion under case headings and permute the columns. Thus the entry would appear as follows:

AGENT	ACTION	OBJECTIVE	TIME	MANNER	LOCATION
	discover	oil			Alaska

and again as:

OBJECTIVE	AGENT	ACTION	TIME	MANNER	LOCATION
oil		discover			Alaska

and so forth.

It appears that this form of index would be easier to use for several reasons. First, it is easier to read than the reverse order articulated index entry with its "dangling preposition". Second, a user normally has very specific requirements in mind when attempting to use an index. Suppose, for example, the entry had been "discovery of oil in the Yukon" rather than "discovery of oil in Alaska." In a standard articulated index, the user would have to scan every "oil, discovery of, ..." entry in order to find this closely related entry. With a tabular approach, once the user located "discover oil" he need only scan the "LOCATION" column to find the related entry.

At least one prototype of this index exists. A mini-abstract produced by Predicasts, Inc. (34) utilizes a tabular approach which roughly corresponds to AGENT-EVENT-OBJECT and to several of the peripheral cases such as time and location. An example is included as Figure 3.9.

Another indexing technique would be to use a reduced portion of the graph as the index entry. Thus the network of Figure 3.7 ("Larry's Trip to Tragedy") might be reduced to the graph of Figure 3.10. This type of entry is not well-suited for printed indexes. It would, however, be a reasonable means of display on a cathode-ray tube. There are several advantages to this method. First, the major "concepts" and the relations between them are clearly delineated. Second, it is relatively easy to distinguish documents. Third, additional detail may be displayed upon user request. For example, the user may be interested in the "time" frame of the incident. This was not a cluster on the graph and is not initially displayed. However, if the entire surrogate were available, the information could be added to the displayed structure. For example, in the "Larry" article, the phrase 'shortly before noon last November 11" would be retrieved as the "time" subgraph and added to the graph. Or, alternatively, the user might ask for more detail concerning a display node. In this case, other relations concerning the node would be displayed.

It appears that the document surrogates might be valuable to all existing indexing techniques. Further the organization of the surrogates suggest some improvements and some innovations in indexing. It is suggested that the surrogates may constitute the "index space" which

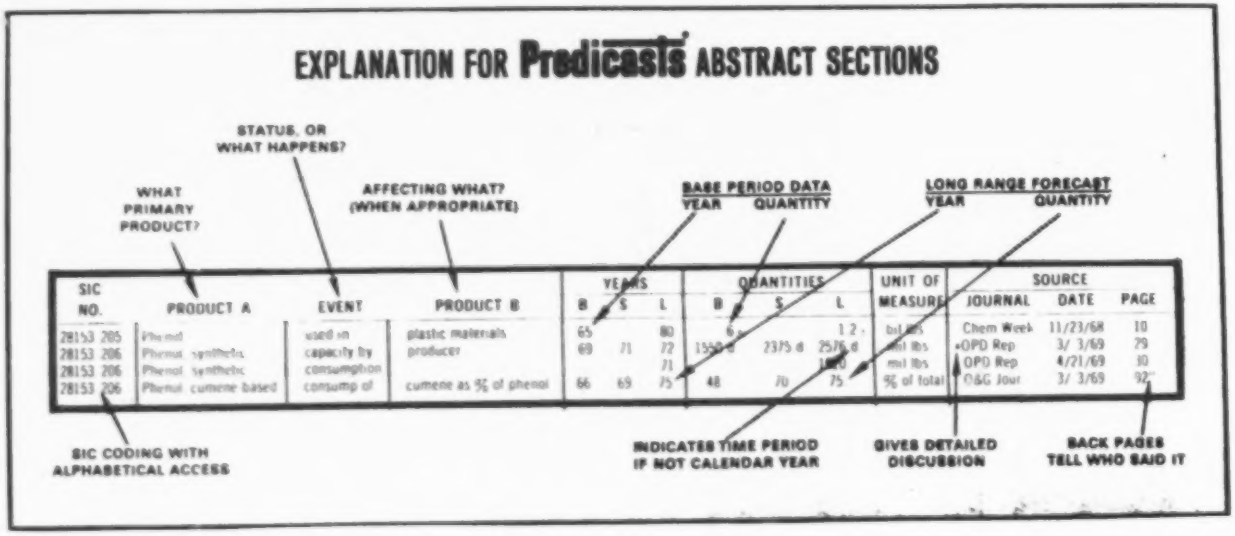

EXPLANATION FOR Predicasts ABSTRACT SECTIONS
STATUS, OR WHAT HAPPENS?
WHAT PRIMARY PRODUCT?
AFFECTING WHAT? (WHEN APPROPRIATE)
BASE PERIOD DATA YEAR QUANTITY
LONG RANGE FORECAST YEAR QUANTITY
SIC NO.
PRODUCT A
EVENT
PRODUCT B
YEARS B S L
QUANTITIES B S L
UNIT OF MEASURE
SOURCE JOURNAL DATE PAGE
28153 205 Phenol used in plastic materials 65 80 12 bil lbs Chem Week 11/23/68 10
28153 206 Phenol synthetic capacity by producer 69 71 72 1550 d 2375 d 2575 d mil lbs OPD Rep 3/ 3/69 79
28153 206 Phenol synthetic consumption 71 1800 mil lbs OPD Rep 4/21/69 10
28153 206 Phenol cumene based consump of cumene as % of phenol 66 69 75 48 70 75 % of total C&G Jour 3/ 3/69 92
SIC CODING WITH ALPHABETICAL ACCESS
INDICATES TIME PERIOD IF NOT CALENDAR YEAR
GIVES DETAILED DISCUSSION
BACK PAGES TELL WHO SAID IT

Figure 3.9 Sample from a Predicasts mini-abstract showing tabular form of display.

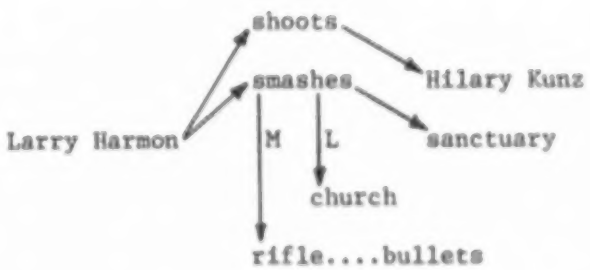


Figure 3.10 Example of an index display generated from the network of Figure 3.7.

Landry and Rush have defined in their work with indexing theory (35).

4.2. Abstracting

But natural language processing involves more than indexing schemes. Can the surrogates be of value in other ways as well?

Consider automatic abstracting. Used as a foundation for present techniques, the surrogates could serve as a basis for a check of the abstract. The abstract network should contain the same clusters in the same relationships as does the parent document. Further, case roles should be consistent.

New abstracting techniques might develop. For example, one might propose that those sentences with multiple clusters be selected to form the abstract. Or a dynamic abstract might be developed. A base structure is displayed to the user who indicates a node. The connections to that node are then displayed and the process is repeated until the user is able to reject or accept the abstract.

A journalistic approach to abstracting might be considered. With the aid of case role assignments, the answers to the questions "who?" "what?" "where?" "when?" "why?" and "how?" could constitute the abstract.

A structural surrogate could free automatic techniques from the necessity of extracting. Nodes and relators could be rearranged to form new sentences.

4.3. Information Retrieval

In the area of question-and-answer systems, several possibilities arise. Since most questions are of the kind mentioned above, and these find correspondence in case frames, a large number of questions might

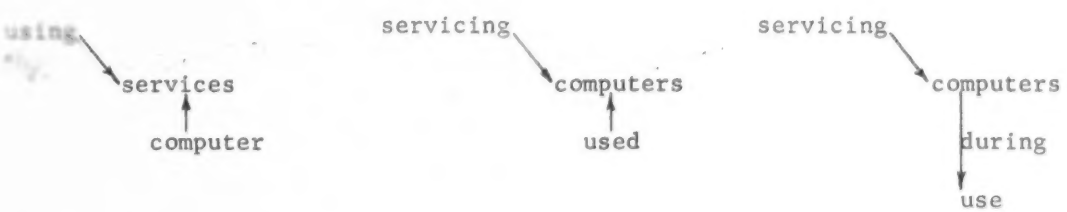
be answered directly from the graph. For example, suppose the question "where has popcorn been found?" were asked. The system retrieves the article on "popcorn", (Figure 3.9) scans for the relator "found" and the locative case frame. The answer is "in Peruvian caves". Suppose, however, the question had been worded "Where was popcorn discovered?" Without "knowing" the relator term, the system might still respond, on the basis of "popcorn" and the locative, with:

1) "Scientists have found popped corn in Peruvian caves."

and

2) "Indians introduced popcorn at Plymouth."

Systems for the selective dissemination of information may also benefit by using structural surrogates. The usual Boolean operators could be replaced with the relational edges of the graph. Thus if the user wished to retrieve "using computer services" (USE + COMPUTER + SERVICE) he would not retrieve "servicing computers during use" (also USE + COMPUTER + SERVICE) or "servicing used computers" (also USE + COMPUTER + SERVICE). The graphs, shown below are clearly not equivalent.



4.4. Other Disciplines

As other disciplines turn to computer assisted processing of natural language to help solve problems in their fields of study, the structural

surrogate base may become even more valuable. Linguists who have shown an interest in the structural patterns across languages would surely benefit from automatic displays of text structures. Those involved with behavorial analysis, e.g., bargaining and arbitration, might find a structured discourse helpful in pinpointing and avoiding "communication breakdowns". Speech patterns may indicate psychological maladjustments and structural surrogates may fit with Pepinsky's schemapiric view of language usage (36). Structural displays may also indicate bias, say in the news media; they may facilitate language education. The possibilities are numerous.

5. Summary

An algorithm (AGNES) has been defined to generate a structural surrogate of English text on the basis of syntactic analysis and case grammar assignments. The feasibility of computerizing this algorithm has been demonstrated. Significant and widespread applications have been suggested for the product of such a program.

What remains to be discussed are the results of this work in light of the criteria set forth in the introduction, and some suggestions for a broader framework in which such a system might be viewed.

CHAPTER IV. RESULTS AND DISCUSSION

> Results! Why, man, I've gotten a lot of results. I know several thousand things that won't work.
>
> T. A. Edison

1. Experimental Results

1.1. The Four Major Trends

Four significant results can be reported as consequences of this research. First, multiple references to the same node, "clusters," appear in the text. Second, the "clusters" generally contain the major "topic" of the text. Third, the case assignments for each noun in a "cluster" are consistent. Fourth, these case assignments generally agree with an "intellectual" analysis of the text.

In the discussion of the results, reference will be made both to the actual graph samples found in Appendix A and the summary of the results presented in Table 4.1.

A cluster is a collection of nodes, the labels of which form an equivalence class. In the three-dimensional model discussed in Chapter III, Section 1.3, the cluster collapses to a single node.

Clusters are evidenced in the two-dimensional graphs by the intersentence relator edge (double dashed lines which run vertically from sentence to sentence). The size of the cluster may be determined by counting the nodes which the intersentence relator edge connects. A network is completely connected if every sentence is linked by at least

Table 4.1 Summary of Case Assignment Distributions

DOCUMENT	CLUSTER	WORD CLASS	TOTAL OCCUR-RENCES	CASE ROLES*									
				MAJOR				MINOR					
				AG	OB	BE	EX	LO	MA	TI	CA	PR	CO
"Larry's Trip to Tragedy"	Larry Harmon	noun	11	9	2								
	church	noun	8		2			6					
	Hilary Kunz	noun	4	1	3								
	shot	verb	4										
	rifle	noun	3		1				2				
	bullets	noun	3						3				
	police	adj	3										
	sanctuary	noun	2		2								
	smashed	verb	2										
"Age-Old Popcorn"	popcorn	noun	5	1	4								
	Indians	noun	2	2									
	popped	adj	2										
"Occurrence of Letters in Engineering Periodical Titles"	titles	noun	10	3	6				1				
	occur-rences	noun	5		3			1	1				
	letters	noun	5		3			1	1				
	counted	verb	4										
	form	noun	4	1	2			1					

*AG-agent, OB-object, BE-beneficiary, EX-experiencer, LO-locative, MA-manner, TI-time, CA-cause, PR-purpose, CO-comitative

Table 4.1 (continuation)

DOCUMENT	CLUSTER	WORD CLASS	TOTAL OCCUR-RENCES	CASE ROLES*									
				MAJOR				MINOR					
				AG	OB	BE	EX	LO	MA	TI	CA	PR	CO
"Is Canada Turning Against Us?"	America Canada	noun	7	5	2								
	U.S.	noun	17	7	2		5	3					
	Canada	noun	29	11	7	1	2	6	1			1	
	trade	noun	1	1									
	trade	adj	4										
	know	verb	5										
	energy	adj	3										
	money	noun	3	3									
	policy	noun	3		2							1	
	dollar	noun	3		1			1	1				
"Automatic Abstracting" (an abstract)	abstracts	noun	4		2			2				1	
	methods	noun	4	1	3								
	abstract	noun	4		2			2					
	auto-matic	adj											
	sentences	noun	2		2								
	system	noun	2		2								
	trend	noun	2	1				1					

*AG-agent, OB-object, BE-beneficiary, EX-experiencer, LO-locative
MA-manner, TI-time, CA-cause, PR-purpose, CO-comitative

one intersentence relator edge to another sentence. The major clusters of each network have been identified in Table 4.1 under the heading "cluster". The number of nodes in the cluster corresponds to the "total occurrences" entry in Table 4.1.

As an example of the extent of the clustering, notice that there is no sentence in the network of "Larry's Trip to Tragedy" (Figure 3.7) which is not connected to at least one other sentence. And in fact, each sentence is in some way finally connected to the "Larry Harmon" node. The article which is graphed in Figure 3.6 ("Age-Old Popcorn") is also completely connected through just three nodes ("popcorn", "popped", and "Indians"). Even the abstract ("Automatic Abstracting and Indexing") Appendix A is completely connected. One might expect the abstract to be less well-connected since the author (in this case a program) is more concerned with brevity than continuity. It can be inferred that this "clustering" is an attribute of most English text. Furthermore, the results clearly indicate that these "clusters" may be automatically discovered.

That the "clusters" generally contain the major "topic" of the text is not as readily inferred. Since it is not always possible for two people to agree upon the "topic" of a text, it is foolish to claim that any procedure could distinguish "the topic" unequivocally. However, if "topic" is roughly defined as the "thing talked about in a text", the results clearly demonstrate that the clusters are in fact "topical". Notice also that the claim is made that the clusters contain the major topic. That is, the set of clusters does not necessarily constitute

the topic, but rather, the topic is a subset of the set of clusters.

Evidence in support of this claim is to be found in Table 4.1. For example, if one were to index "Larry's Trip to Tragedy" using single terms, the terms chosen would probably be included in the list,

Larry Harmon
church
Kunz
rifle
sanctuary
bullets
police
shot
smashed

Likewise the article "Is Canada Turning Against Us?" would almost certainly be indexed under "Canada", "U.S." and possibly under several of the minor clusters "trade", "dollar", "policy", etc., because the article is "about" U.S.-Canadian relations in the light of such issues as trade balance, the energy crisis, and foreign policy. The research to date indicates (as one might reasonably expect) that the major topics of a text are a subset of the "clusters".

The third claim is that the case assignments for all the nouns in a cluster generally agree. This result can best be demonstrated by reference again to Table 4.1, where case assignments have been summarized. In general, the case assignments are not scattered throughout the categories but are concentrated in one case assignment per cluster. In "Larry's Trip to Tragedy" for example, there are 11 nodes of the "Larry Harmon" cluster, 9 of which are agentive. Of the 8 nodes of the "church" cluster, 6 are locative. All the nodes of the "bullets" cluster have the manner case.

An apparent exception to the consistency of case assignment is the article "Is Canada Turning Against Us?" Here the tallies for several of the clusters are spread across many case assignments. Notice, however, that the major case assignments of "U.S." and "Canada" far outweigh the minor. Furthermore, the other terms are more strongly minor than major. Thus, even in this "exception", the major case assignments are discernible. In general, the "case consistency" of a cluster has been demonstrated. The significance of this result is that, assuming these case assignments are "correct", not only is the topic of a text a subset of the clusters, it is an identifiable subset which can be meaningfully labeled. The assumption of "correctness" is defended next.

The last claim, that case assignments agree with an "intellectual" analysis of the text, is problematic because of the term "intellectual". Intellectual analysis refers to the manual procedure of determining the relevance and role of each cluster. What is meant by this statement is that if an article is "about", for instance, Curie's discovery of radium, and the clusters are "Curie" and "radium", then to be "correct" the "Curie" cluster should be primarily agentive and the "radium" cluster should be primarily objective.

A study of the graphs (Figures 3.6 and 3.7 and Appendix A) and Table 4.1 reveals the claim is well-founded. "Larry's Trip to Tragedy" is "about" a young man named Larry Harmon who enters a church with sledgehammer and rifle, shoots and kills a sacristan named Kunz and wrecks the sanctuary. He leaves the building and shoots several other

people before he is shot and killed by police. "Larry Harmon" is clearly the "agent". The location of the incident is a "church". Possible objects are "Kunz" and "sanctuary" and candidates for the manner role are "bullets" and "rifle". To summarize now from the case roles, a "Larry Harmon" did something to a "Kunz" and a "sanctuary" with "bullets" and "rifle" at a "church".[3] The case role analysis is remarkably similar to the "intellectual" synopsis.

If there are multiple agents in an article, the case roles clearly indicate both, as in the "tie" of "U.S." and "Canada" in "Is Canada Turning Against Us?". One thing which is not clear from the case role summary is whether the article is partly about the U.S. and partly about Canada or if it is about their interactions. The answer is evident if one notices that in the graph, the relator edges "criss-cross" from subject to object to subject again. This illustration serves as a reminder that the surrogate is not merely a means to an end; not a process, but a product.

One problem with the analysis of "Is Canada Turning Against Us?" is its failure to indicate the agent-object balance which is found in the article. Upon examination of the text, it was found that the author used adjectival forms more often in the object than in the agent slot. For example, "Still, many Canadians look upon U.S. holdings...". It is suggested that perhaps adjective clusters should also be

---

3. This sentence is formed by a reversal of the rules which define a case assignment. For example "at" denotes a locative; thus the locative is made object of the preposition "at".

classified, perhaps by the case roles of the words they modify. Whether the case assignments would remain consistent has not yet been determined. An alternate analysis of this article is displayed in Table 4.2. One might further hypothesize that classification of verbs (stative, process, action, action-process) would be beneficial. For example, the statement "Larry took aim...and shot Kunz" might be tied to "Clark...was shot", "Brass...was shot" and "Schroeder...was shot" to produce "Larry shot Kunz, Clark, Brass and Schroeder." At this point, however, such transformations are purely speculative.

In brief, experimentation has demonstrated that "topical, case-consistent clusters" which adequately "describe" an English text may be automatically obtained, related and labelled to form a coherent synopsis of that text.

1.2. Observed Trends

Several other trends have been observed which, though not fully demonstrated in the samples, warrant mention.

For example, of the 18 action and action-process verbs in "Larry's Trip to Tragedy", "Larry Harmon" nodes are agent of all but two. In other words, not only is "Larry Harmon" the major agent, he is virtually the only one.

Second, as has been suggested, adjectival roles may assist in the total description of an article.

Third, case assignments, even in isolation, may provide valuable information in the comparison of documents. For example "Age-Old Popcorn" is clearly about popcorn and no strong relations to other

Table 4.2 Revised Analysis Using Adjectival Cases

DOCUMENT	CLUSTER	TOTAL OCCURRENCES	CASE ROLES*									
			MAJOR				MINOR					
			AG	OB	BE	EX	LO	MA	TI	CA	PR	CO
"Is Canada Turning Against Us?"	America-Canada	17	7	6		2	2					
	Canada	49	13	17	4	6	7	1			1	
	U.S.	42	13	13		5	8	2				1

*AG-agent, OB-object, BE-beneficiary, EX-experiencer, LO-locative
MA-manner, TI-time, CA-cause, PR-purpose, CO-comitative

clusters exist. However, the objective case may distinguish this article from another about popcorn.

Fourth, it appears that the "subject" of an article is the largest cluster of those bearing major case roles. For example, in "Larry's Trip to Tragedy", the two largest clusters are "Larry Harmon" and "church", the former being largely agentive and the latter locative. "Larry Harmon" is clearly the subject. In "Age-Old Popcorn", "popcorn" and "Indians" both take major case roles, but "popcorn" is the larger cluster and is clearly the "subject" (though not the "agent") of the article. Further research is necessary to verify these observations.

However interesting the results may seem, they are incidental if viewed in isolation. The world is not interested in a "good" mousetrap but in a "better" one. How does the surrogate approach compare with existing methods of document representation?

2. Comparison with Existing Systems

A comparison of document representations in themselves is difficult. Rather the possible products derived from the surrogate will be compared with products which now exist. The discussion will be limited to indexing procedures for the purpose of brevity.

Assume that a keyword or uniterm index is desired. If the index is to be produced automatically, its production probably depends upon statistical procedures. In "Larry's Trip to Tragedy", assuming a statistical program were able to associate "Larry", "Harmon", and "Larry Harmon", it would find 6 occurrences of "Larry". However, it

would find 8 occurrences of the word "church". Suppose that only one index term is desired. Then "church" would be chosen, because it appears more often. If the surrogate were used, on the other hand, the largest cluster would be "Larry Harmon" since equivalence relationships are identified. Furthermore, the case role assignment identifies "Larry Harmon" as a major case and "church" as a minor case. The entry would be "Larry Harmon", a better choice. Similarly, in "Age-Old Popcorn", "popcorn" appears twice, as does "corn", "popped" and "Indian". A statistical method cannot distinguish between adjective and noun, nor can it recognize equivalence relationships. "Popped" is as likely a choice as "popcorn" or "Indian". The surrogate, however, equates "popcorn", "corn" and "it" for a major-case cluster of "popcorn" and correctly chooses the proper term.

If a KWIC index is to be produced, the title is usually used. This approach works well for "Age-Old Popcorn" since "popcorn" is probably sufficiently specific. "Larry's Trip to Tragedy" is actually a longer article than is graphed, and deals ultimately with the question of LSD and drug abuse. A KWIC index of the title would lose (except in the idiomatic use of "trip") any reference to drugs. "Is Canada Turning Against Us?" survives the KWIC fairly well until one realizes that "Us" has little value without knowledge of the place of publication. The only remaining "meaningful" term is "Canada". If one assumes that an abstract is entitled as is its parent document, the abstract (Appendix A) also KWICs well. However, if the title of the abstract is "Abstract," no information is gained.

An alternative to using the title as the basis for KWIC index entries is suggested by the surrogate structure. A sentence such as "Larry Harmon ⟶ Hilary Kunz and sanctuary with bullets and rifle at church" (which is again produced by a reversal of the case assignment rules), produces a better basis. In the article "Is Canada Turning Against Us?" the surrogate clearly identifies "us" as the "U.S." and in the extended analysis provides a reasonable description of the text.

Manual procedures to produce an articulated index could be improved or replaced with automatic procedures using the structural surrogate. Rather than reading the entire text to extract an articulated entry, an indexer could merely analyze the graph to determine the articulation phrases. For example, the indexer might be presented with the graph in Figure 3.10 from which to construct the entry. This is certainly easier than reading and analyzing four paragraphs of text.

Several other examples of automatic indexing procedures using the surrogates are described in Chapter III, Section 4. These methods may be as good as or better than existing manual techniques and better and more versatile than existing automatic techniques.

The structural surrogates exhibit certain interesting characteristics which appear to justify their use in natural language processing. However, their "performance" must be judged ultimately in the light of their design criteria.

3. Comparison with Design Criteria

It was proposed that the syntactic structural surrogates would be a representation of any English text more suitably organized than a linear string. These criteria have been met. The organization of the surrogate is based on syntactic and equivalence relationships. And since the algorithm does not rely on analysis documents, such as dictionaries, it is applicable to any English text.

The next criterion is that the surrogates may be produced by computer. This criterion has not been met explicitly, but the feasibility of computerization has been discussed in Chapter III, Section 3.

The next criterion is that the major concepts of the text be discernible or derivable from the surrogate. This criterion is met in that document "clusters" which correspond to its "topics" are automatically derivable. Case assignments further determine the roles of the clusters and the graph illustrates the syntactic relationships between clusters.

In the sense that the criterion implies that the surrogate system can replace a trained indexer, the criterion has not yet been met. A surrogate system, like the indexer, must be trained. Questions such as "how many nodes determine a 'large' cluster?" and "if a decision is to be made, shall the basis be size of cluster or case role?" need to be answered before automatic processing can be accomplished. Yet the concepts of a text (though perhaps not the index terms) are "clearly discernible" and the criterion is thus met.

The next criterion is that the document must be derivable from the surrogate. There are two ways of viewing this demand. If one were to require that a user to able to reconstruct the original text given the surrogate, the criterion has not been met. No means are included to indicate the order of the words in a sentence or the sentences in a text. At least the determiners, and in some cases stative verbs and pronouns, are eliminated in the graph. In some instances case markers replace prepositions. If however, one may assume that if the document is to be derived from the computer representation of the surrogate, then it is quite possible that the ordering of words and sentences and the occurrence of all words and punctuation could be preserved. In that case, the document could easily be reproduced and the criterion would be met.

The last criterion is that the surrogate be economically feasible to implement. It is difficult to measure system "performance" in this respect because the IBM S/370 configuration at Ohio State University does not include a graphics terminal, and it is generally difficult to judge economics with a limited sample size. The Young programs process at least 7,700 words per minute in 252 K of main storage. It is not possible to estimate the additional storage and time requirements for the graphic programs if it were to be programmed on the S/370. It is clear that no human preprocessing, processing or interpretation need enter the procedure. The criterion is not met, but neither is it ruled out. It is simply not possible to judge.

Of the six design criteria, four have been met, one has been partially satisfied and one remains to be tested. The "performance" of the surrogates in this respect is quite good.

4. A Word on Accuracy

It is customary to include in a discussion of experimental results, mention of such things as accuracy, efficiency, experimental error, etc. An estimate of accuracy for the algorithm is difficult to obtain. In fact accuracy for an algorithm of this type is difficult to define. It is assumed that accuracy must in some way be bound by the accuracy of the inputs. Young claims an average accuracy for the case grammar analysis of 73% (a figure that reflects the cumulative effect of errors produced in each of the preceding analysis steps). Since AGNES makes no "assignments" per se, but maps assignments onto a graph, the only "errors" the program makes are (a) those involving sentences which are wrong in the case grammar analysis or (b) those involving sentences for which the rules are incomplete and a graph cannot be drawn. One would expect that type B errors could be eliminated. Thus an estimate of an accuracy of about 70% for the system overall is reasonable. However, the surrogate is one case in which the whole seems to be greater than the sum of its parts. That is, even if several sentences were erroneously graphed, there is a good probability that the resultant graph would still adequately describe the text. For example, suppose the second sentence of "Age-Old Popcorn" were in error so that no relation between popcorn and corn were made. The graph still "clusters" at "popcorn",

"corn" and "Indian" and at worst, a dual object of "corn" and "popcorn" is assumed. More often, a missing or erroneous sentence would make little difference in the surrogate and no difference in the end product, the index entry.

Any scientific research is a cyclic and dynamic process. A hypothesis-test evaluation is followed by revision and retesting of the hypotheses. There is never really a conclusion, per se, to the research. At this point, however, it would be good to summarize the present endeavor; look at this segment of research in a broader perspective and suggest what direction further study of the topic might take.

CHAPTER V. SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

> Grant, O God, That we may always be right, for Thou knowest we will never change our minds.
>
> Old Scottish Prayer

1. Summary

It was suggested in the Introduction to this thesis that an adequate representation of language, at least of written English, is both crucial to and lacking in computer-based language processing. A graphic surrogate of written English has been proposed, defined and illustrated which the author believes closely approximates the required representation of text. The surrogate makes explicit three important properties of language: context, syntactic function and case role.

Certainly the contextual properties of the surrogate differ markedly from those of the linear strings (sentences) from which they are derived. The importance of this observation is that the relationships between the elements of text are made explicit and readily discernible in the surrogate. And the surrogate represents the conviction that language is multi-dimensional rather than one-dimensional (as would appear to be the case if one takes written or spoken language at face value).

The construction and organization of the surrogates is syntactically based. Thus the shape of a graph is determined by the syntax of the sentence. The values associated with its nodes and edges are determined by the vocabulary.

Whereas syntax defines relations among the elements of text, case roles characterize those relations. Such characterization, in turn, makes possible automatic "judgements" concerning specific elements of the text.

An algorithm has been developed which generates the structural surrogates using the results of a syntactic analysis system described elsewhere (7). An important feature of all the procedures involved in the production of the surrogates is that they are independent of subject area, they are efficient, and they produce quite accurate overall results. The generation of the structural surrogates relies solely on the output of the syntactic analysis system and upon a set of rules which prescribe the type of edge that links nodes of the structure.

A small set of rules for construction of the surrogates has been described and several extensions which expand their utility are presented. The algorithm which produces the graphic display is simple and efficient.

Preliminary application of the algorithm (AGNES) to a variety of texts has yielded promising results. Although experimental results are not extensive, the value of the structural surrogates as representations of English text appears to surpass that of linear strings perhaps in all respects save human output.

2. The Surrogates in Perspective

Pepinsky (36) considers three levels in the study of language which he calls, collectively, the "schemapiric view" of language study. The

three levels are the empirical, the analytical and the formal (see Figure 5.1). For our purposes, the written text lies at the empirical level; that is, the string of alphabetical symbols which forms the text is taken as the starting point--one might call it the observational level. At the analytical level are all those procedures which impose some structure on the empirical data. In this case, the syntactic analysis procedures of Young (7) constitute the analytical level in this research. At this level, word boundaries, word classes, phrase boundaries, phrase types, clause boundaries, clause types and finally case roles are ascribed to the text.

At the formal level, which might be called the synthetic level, lies the structural surrogate. The formal level may be considered a "metalanguage", the "language" one uses to talk about languages. The structural surrogate corresponds with the empirical data through a synthesis of the various structures imputed to the empirical data by the analytic procedures.

Once at the formal level, one may then take the data there to be the empirical data for a new series of investigations.

3. Directions for Future Research

> The world is round and the place which may seem like the end may also be only the beginning.
>
> Ivy Baker Priest

It is characteristic of scientific endeavor that research ultimately asks more questions than it answers. The present work offers no exception.

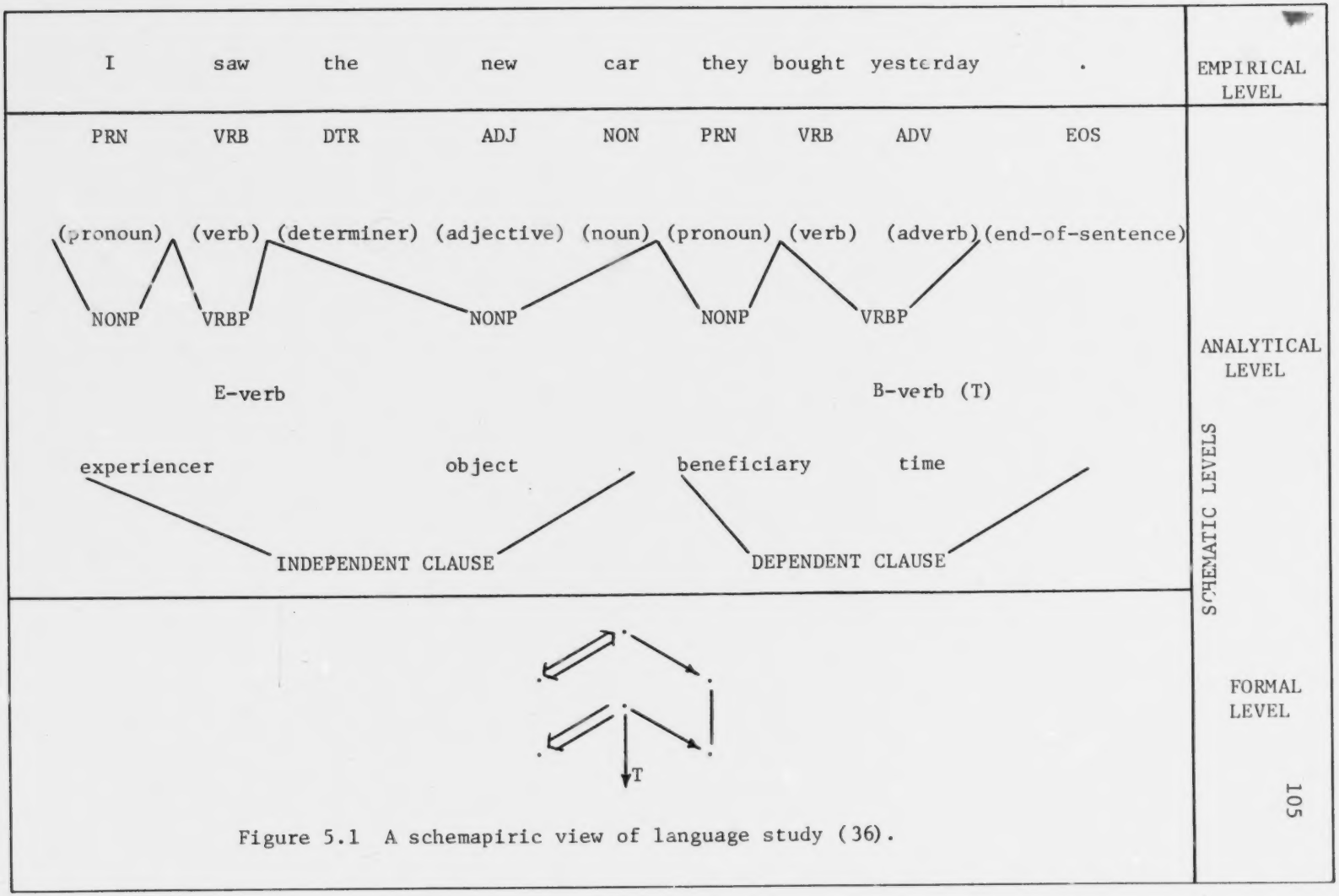


Figure 5.1 A schemapiric view of language study (36).

Further work in the area of this research might proceed in two general directions: the refinement and evaluation of the present system and an investigation into applications of the structural surrogates.

Work on the present system might include the following tasks. An operating graphics program should be implemented and the algorithm evaluated on the basis of its output. An extensive study of pronominal antecedents and other intersentence relators is in order. Further reductions or amplifications of the graphic display should be considered. Questions pertaining to the determination of concept clusters need to be answered. For instance, the questions "What is the relative importance of case, absolute cluster size, relative cluster size?" and "Should adjectives and verbs be classified?" are important to the extension of the work. The system might be extended to other languages. Economic feasibility must be evaluated.

Possible research topics in applications include investigation of the various natural language processing techniques suggested earlier in this paper: the extended articulated index, graphic index entries, dynamic on-line indexes, dynamic on-line abstracts, "journalistic" abstracts, question-answer systems based on case grammar, relational edge "operators" to replace Boolean operators. An extensive comparative study is in order. Stylistic analysis by structure and case comparisons could be explored, as well as the elusive machine translation.

Application of the structural surrogate may be made to any discipline in which natural language or communication is of some importance. Possibilities are endless. What has been developed is a

groundwork for interdisciplinary investigation of communication problems.

REFERENCES

1. F. B. Libaw, "A New Generalized Model for Information Transfer: A Systems Approach", American Documentation 20(4), 381-384 (1969).

2. M. R. Quillian, "The Teachable Language Comprehender: A Simulation Program and Theory of Language", Communications of The Association for Computing Machinery 12(8), 459-276 (1969).

3. L. B. Doyle, "Semantic Roadmaps for Literature Searches", Journal of the Association for Computing Machinery 8(4), 553-578 (1961).

4. D. C. Clark and R. E. Wall, "An Economical Program for Limited Parsing of English", AFIPS Conference Proceedings 27 (Part 1) 307-316 (1965).

5. S. Marvin, J. Rush, C. Young, "Grammatical Class Assignment Based on Function Words", Seventh Annual National Colloquium on Information Retrieval, Philadelphia, Pennsylvania, May, 1970.

6. J. Thorne, P. Bratley and H. Dewar, "The Syntactic Analysis of English by Machine", in D. Michie (ed.), Machine Intelligence 3, American Elsevier Publishing Co., New York, New York, 1968, 281-310.

7. C. Young, Design and Implementation of Language Analysis Procedures with Application to Automatic Indexing, Ph.D. dissertation, The Ohio State University, 1973.

8. C. Young, Grammatical Assignment Based on Function Words, M.S. thesis, The Ohio State University, 1970.

9. C. L. Bernier and K. F. Heumann, "Correlative Indexes III: Semantic Relations Among Semantemes--The Technical Thesaurus", American Documentation 8(1), 211-220 (1957).

10. R. Fugmann, H. Nickelsen, I. Nickelsen and J. Winters, "TOSAR--A Topological Method for the Representation of Synthetic and Analytical Relations of Concepts", Angewandte Chemie, International Edition (in English) 9(8), 589-595 (1970).

11. Ibid., p. 593.

12. R. C. Shank and L. Tesler, "A Conceptual Dependency Parser for Natural Language", International Conference on Computational Linguistics, Stockholm, Sweden, Preprint No. 2 (1969).

13. D. G. Bobrow, "Syntactic Analysis of English by Computer--A Survey", AFIPS Conference Proceedings 24, 376 (1963).

14. S. Klein and R. Simmons, "A Computational Approach to Grammatical Coding of English Words", Journal of the Association for Computing Machinery 10(3), 334-347 (1963).

15. Ibid., p. 338.

16. N. Chomsky, Aspects of the Theory of Syntax, M.I.T. Press, Cambridge, Massachusetts, 1965.

17. W. A. Woods, "Transition Network Grammars for Natural Language Analysis", Communications of the Association for Computing Machinery 13(10) 591-606 (1970).

18. D. B. Vigor, D. Urquhart and A. Wilkinson, in B. Meltzer and D. Michie (ed.), Machine Intelligence 4, American Elsevier Publishing Co., New York, New York, 1969, 271-284.

19. D. G. Hays, "Dependency Theory: A Formalism and Some Observations", Language 40(4), 511-525 (1964).

20. T. Winograd, "Understanding Natural Language", Cognitive Psychology 3, 1-191 (1972).

21. M. A. K. Halliday, "Notes on Transitivity and Theme in English; Part 1", Journal of Linguistics 3(1), 37-87 (1967).

22. C. C. Fries, The Structure of English, Harcourt, Brace & World, Inc., New York, New York, 1952.

23. W. A. Cook, S.J., personal communication, 1972.

24. G. Salton, "Manipulation of Trees in Information Retrieval", Communications of the Association for Computing Machinery 5(2), 105-114 (1962).

25. C. J. Fillmore, "The Case for Case", in E. Bach and R. Harms, (eds.), Universals in Linguistic Theory, Holt, Rinehart & Winston, New York, New York, 1968, 1-88.

26. W. L. Chafe, Meaning and the Structure of Language, University of Chicago Press, Chicago, Ill., 1970.

27. Ibid., p. 97.

28. J. Anderson, The Grammar of Case: Towards a Localistic Theory, Cambridge University Press, Cambridge, England, 1971.

29. C. A. Montgomery, "Linguistics and Information Science", Journal of the American Society for Information Science 23(3), 195-218 (1972).

30. Ibid., p. 197.

31. W. A. Cook, S.J., "A Case Grammar Matrix", Languages and Linguistics Working Papers No. 6, Georgetown University Press, Washington,D.C., 1972, 15-48.

32. J. M. Walsh and A. K. Walsh, Plain English Handbook, McCormick-Mathers, Inc., Columbus, Ohio, 1959, p. 22.

33. F. T. Parsons, How to Know the Ferns, Dover Publications, New York, New York, 1961, 77.

34. R. E. Maizell, J. F. Smith, and T. E. R. Singer, Abstracting Scientific and Technical Literature, Wiley-Interscience, New York, New York, 1971, 74.

35. B. C. Landry and J. E. Rush, "Toward a Theory of Indexing -II", Journal of the American Society for Information Science 21(5), 358-367 (1970).

36. H. B. Pepinsky, "A Metalanguage for Systematic Research on Human Communication Via Natural Language", submitted for publication in the Journal of the American Society for Information Science.

REFERENCES FOR DATA SOURCE

d1. "Age-Old Popcorn," Chemistry, 46 (1), 4 (1973).

d2. J. Reddy, "Larry's Trip to Tragedy," Reader's Digest 100 (601), 120-122 (1972).

d3. A. N. Yerkey, "Occurrence of Letters in Engineering Periodical Titles," Journal of the American Society for Information Science, 22 (4), 290-292 (1971).

d4. L. T. Merchant, "Is Canada Turning Against Us?" Reader's Digest 100 (601), 138-141 (1972).

d5. J. E. Rush, R. Salvador and A. Zamora, "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria," Journal of the American Society for Information Science, 22 (4), 260 (1971).

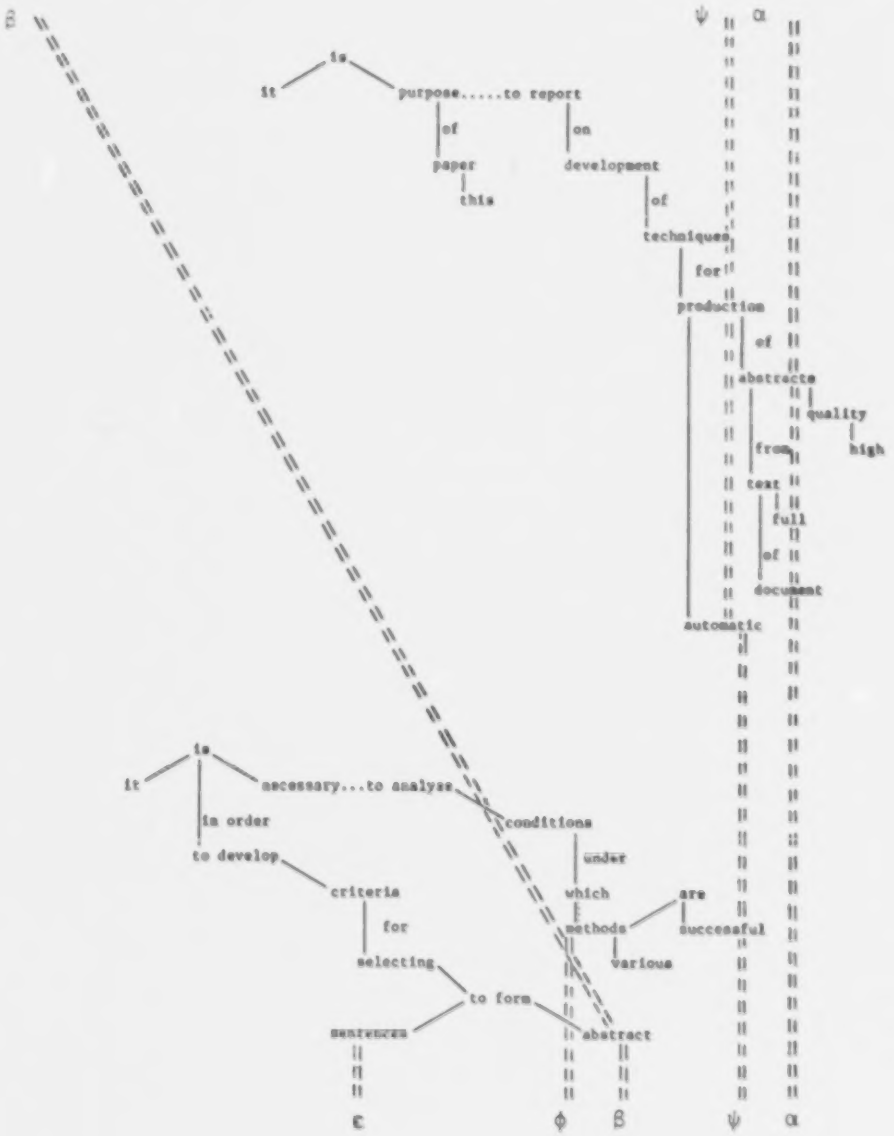
APPENDIX A

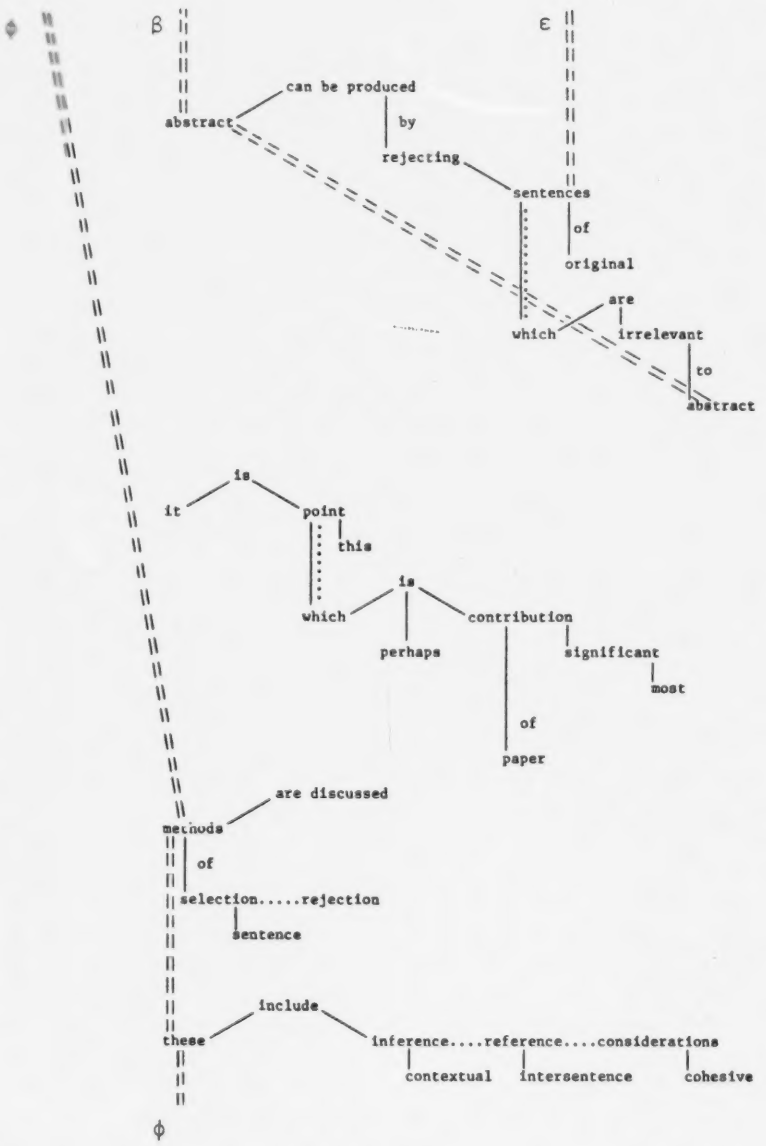
APPENDIX A: SAMPLE AGNES GRAPHS

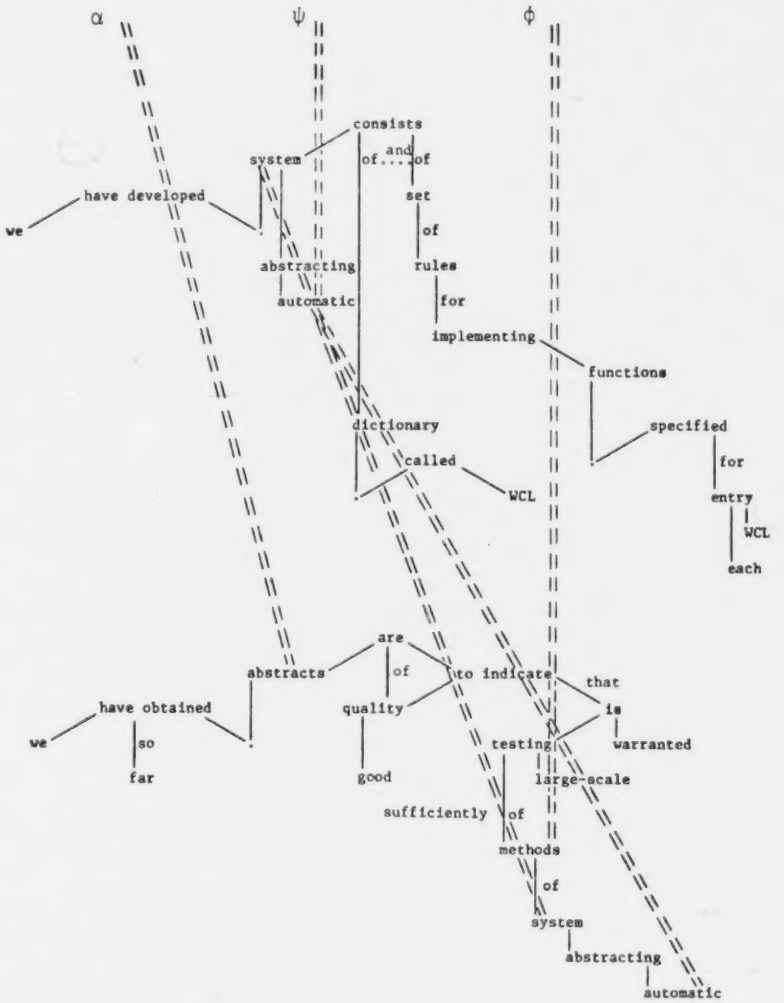
The graphs which follow have been generated using the basic AGNES algorithm and the intersentence relator rules. Included in the sample presented are an abstract, a short technical article, and a portion of a general interest magazine article.

1. AGNES graph of the abstract of "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria" (d5).

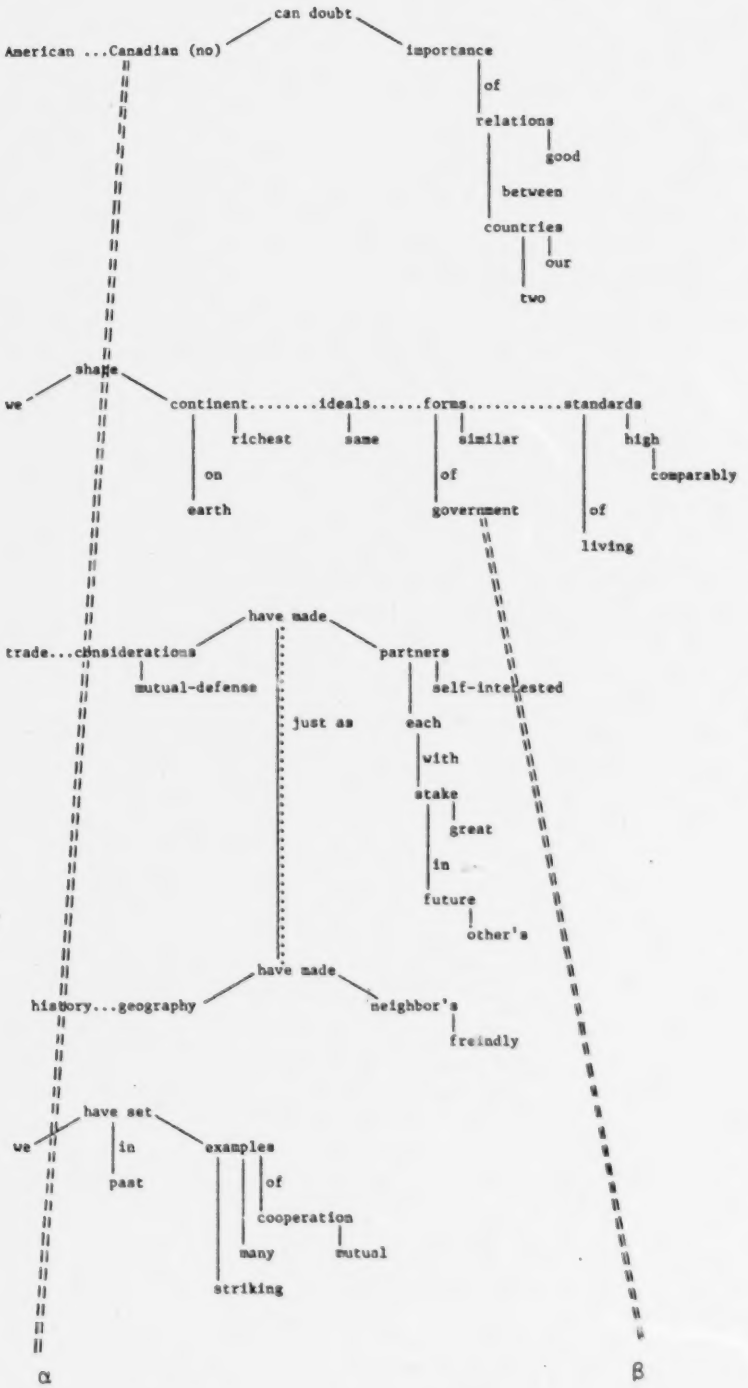
have lent
factors
impetus
of
to
time..cost..value
trend
together with
toward
shortage
generation
of
of
abstractors
abstracts...indexes
qualified
automatic
increasing
has caused
trend
to be placed
this
emphasis
on
increased
abstract
as
locus
of
data
for
systems
retreival
automatic
necessitates
this
creation
of
abstracts
quality
high
α
β
ψ

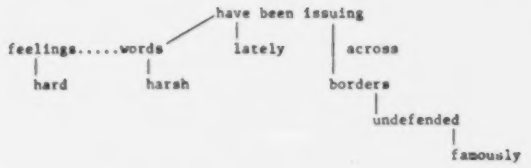

it
is
purpose.....to report
of
paper
this
on
development
of
techniques
for
production
of
abstracts
quality
high
from
text
full
of
document
automatic
it
is
necessary...to analyse
conditions
in order
to develop
criteria
for
selecting
to form
sentences
abstract
under
which
are
methods
successful
various
ε
φ
β
ψ
α

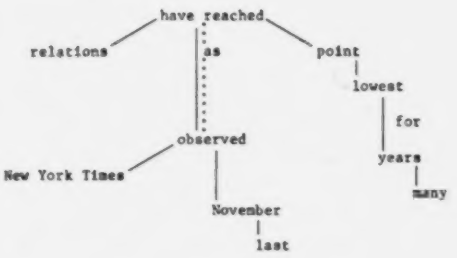

Φ
β
ε
abstract
can be produced
by
rejecting
sentences
of
original
which
are
irrelevant
to
abstract
it
is
point
this
which
is
perhaps
contribution
significant
most
of
paper
methods
are discussed
of
selection.....rejection
sentence
these
include
inference....reference....considerations
contextual
intersentence
cohesive
Φ

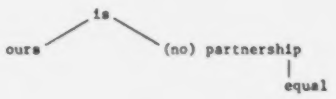

α
ψ
φ
we
have developed
system
consists
of...and...of
set
of
rules
for
implementing
functions
specified
for
entry
WCL
each
abstracting
automatic
dictionary
called
WCL
we
have obtained
so
far
abstracts
are
of
quality
good
to indicate
that
is
warranted
testing
large-scale
sufficiently
of
methods
of
system
abstracting
automatic

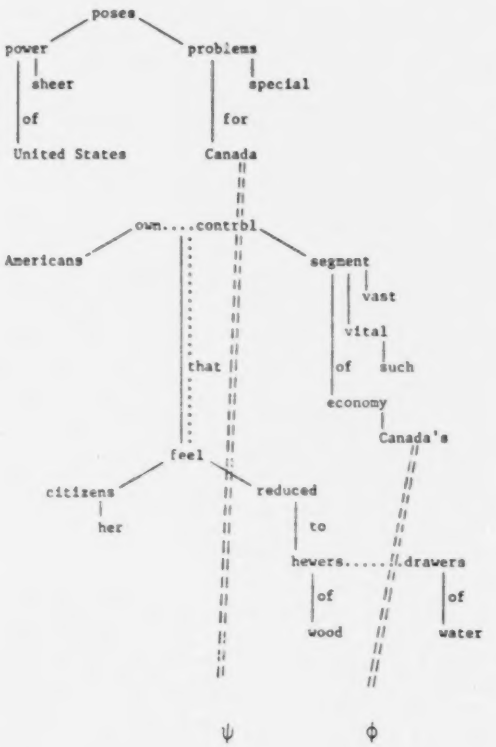
2. AGNES graph of a portion of a general interest article, "Is Canada Turning Against Us?" (d4). The entire article has been summarized in Tables 4.1 and 4.2, although only a portion appears here.

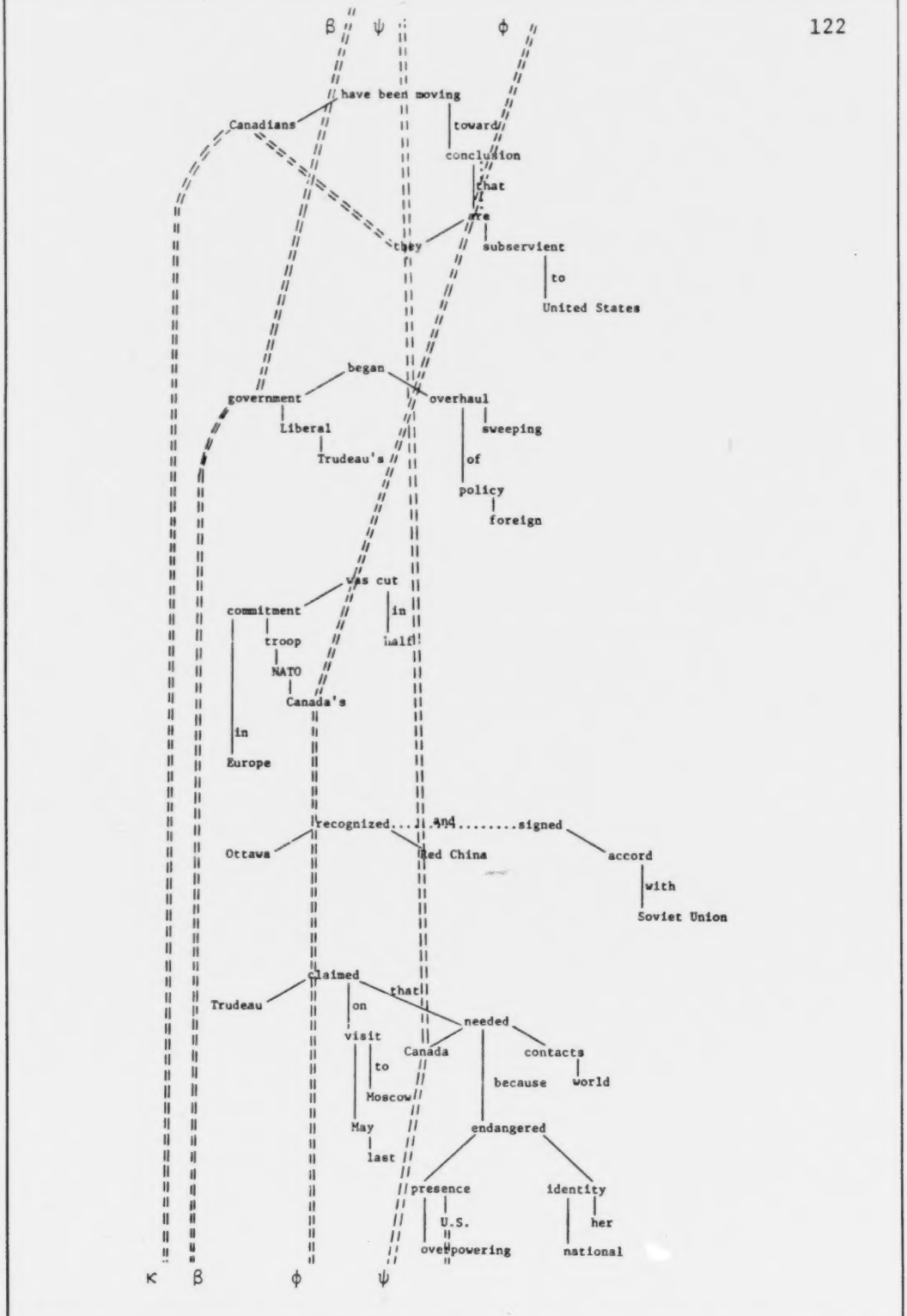

can doubt
American ...Canadian (no)
importance
of
relations
good
between
countries
our
two
shape
we
continent.....ideals.....forms.....standards
richest
same
similar
high
comparably
on
earth
of
government
of
living
have made
trade...considerations
partners
mutual-defense
self-interested
just as
each
with
stake
great
in
future
other's
have made
history...geography
neighbor's
freindly
have set
we
in
past
examples
of
cooperation
many
mutual
striking
α
β

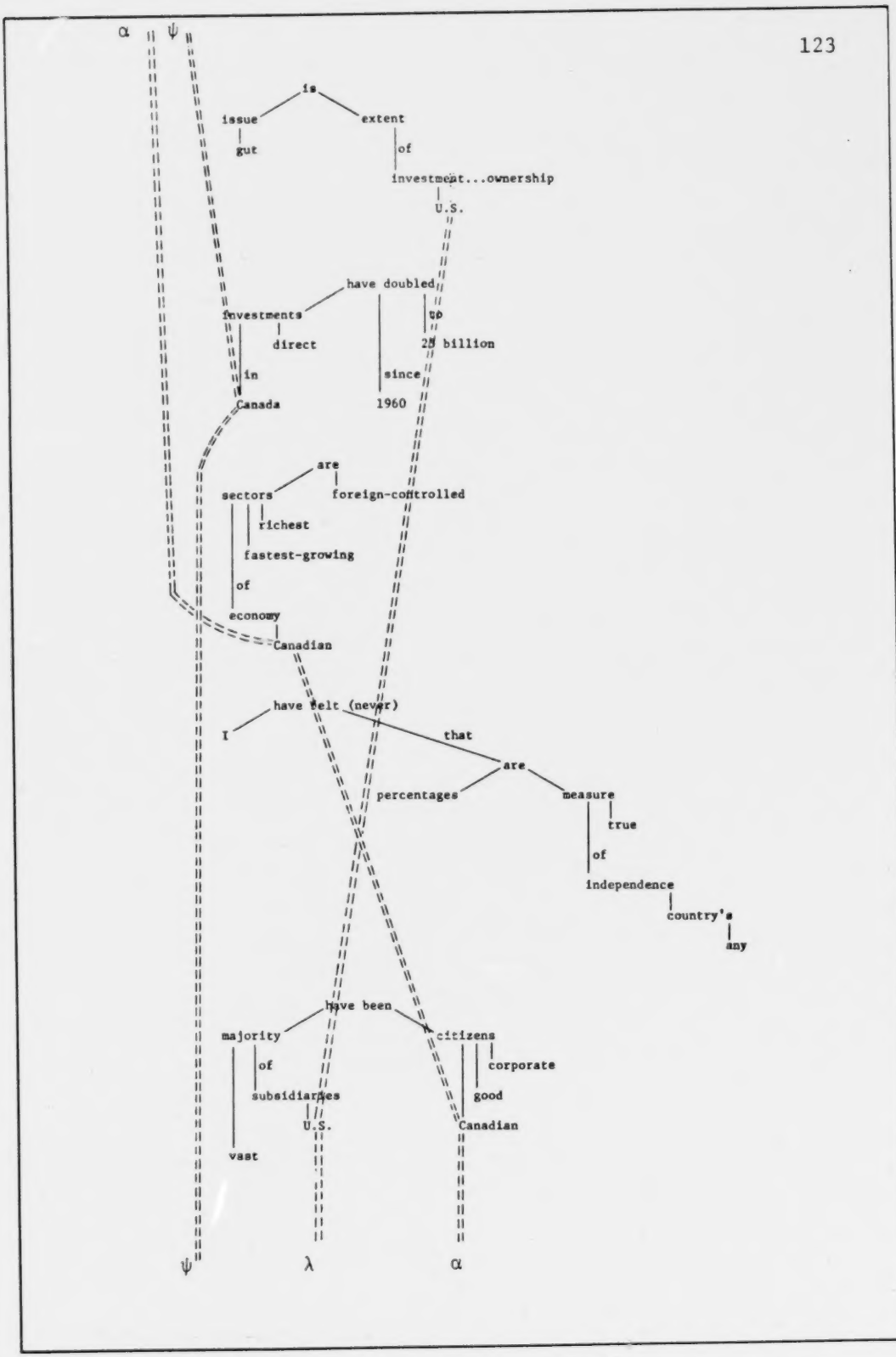

have been issuing
feelings.....words
lately
across
hard
harsh
borders
undefended
famously

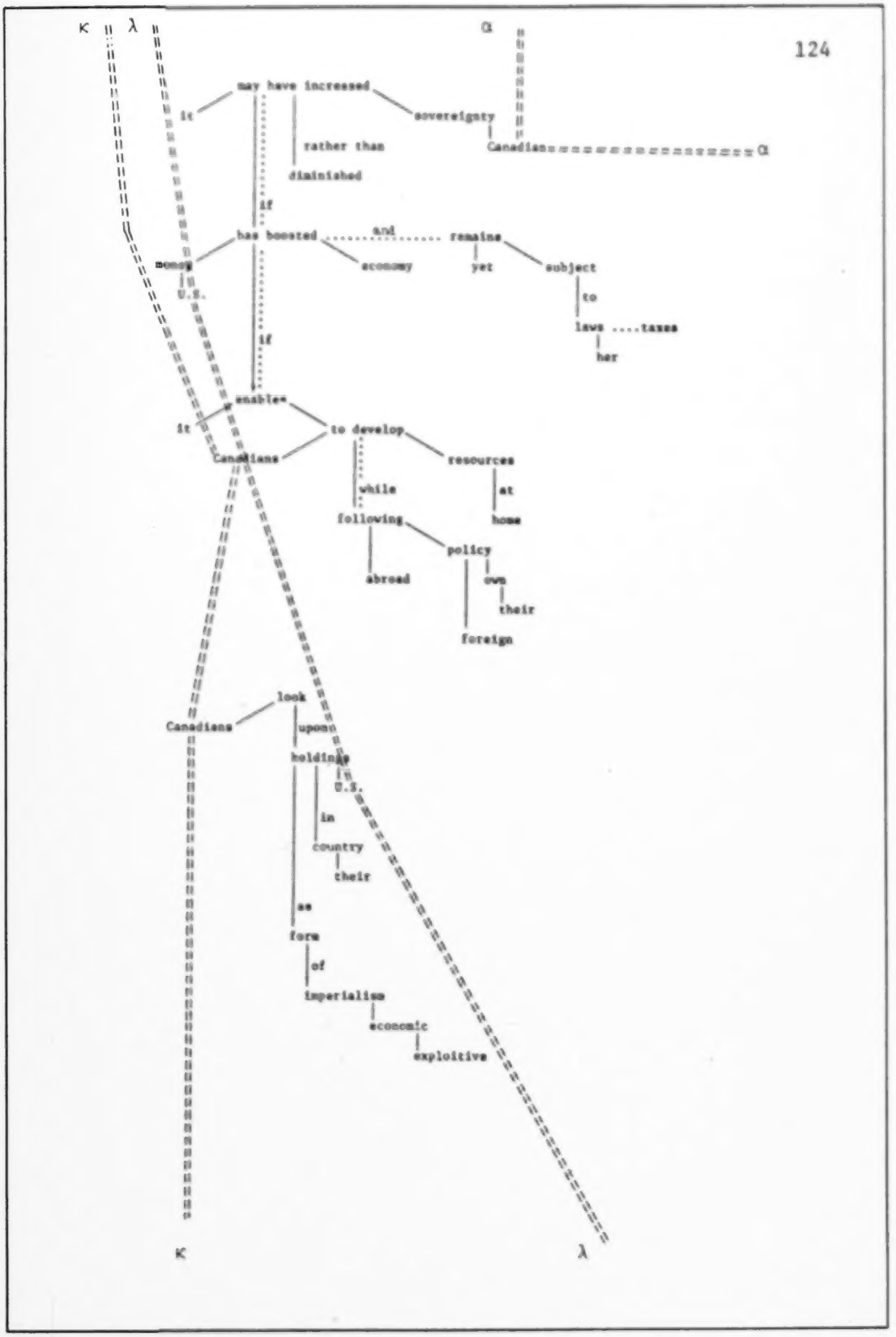

have reached
relations
as
point
lowest
for
observed
years
many
New York Times
November
last

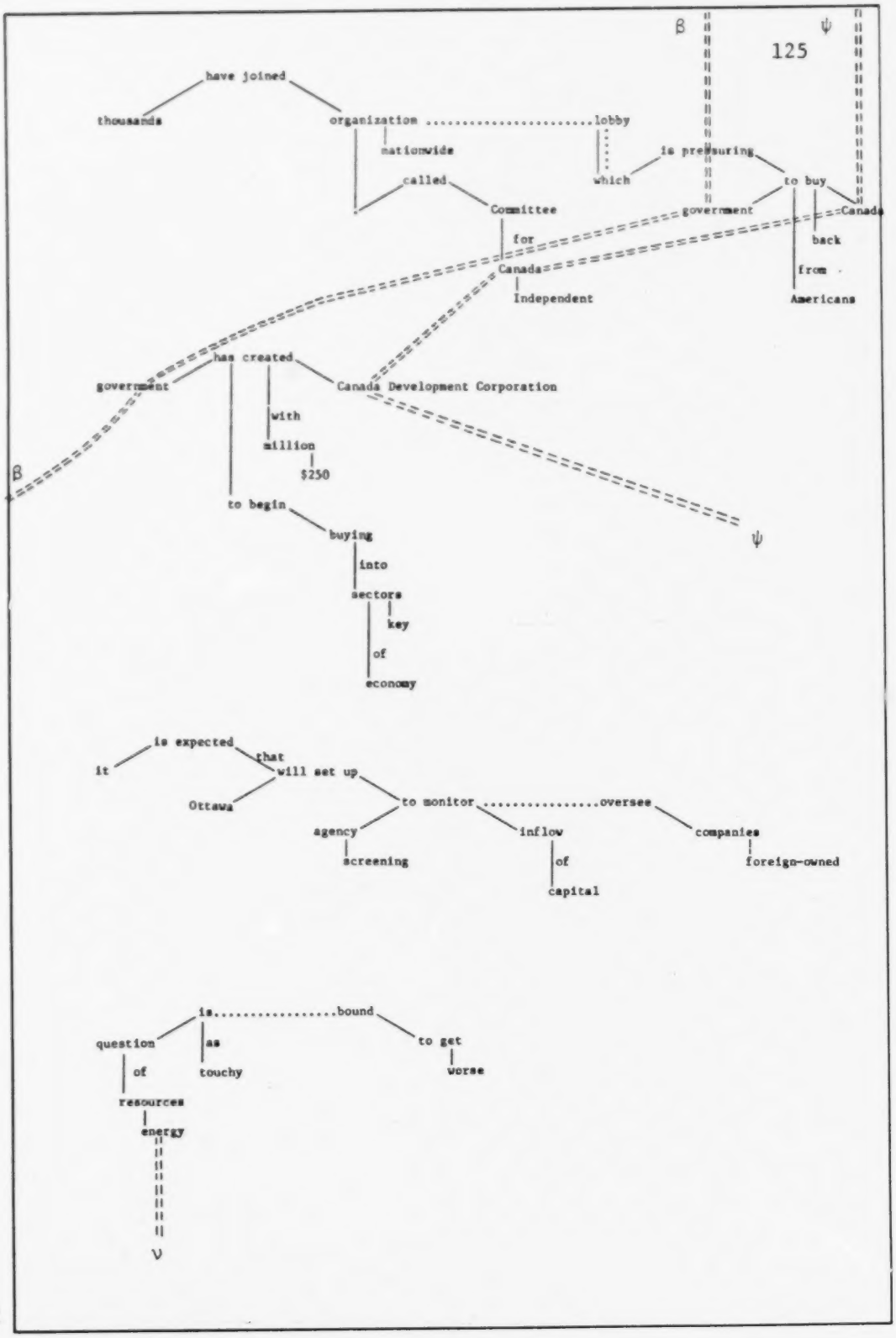

is
ours
(no) partnership
equal

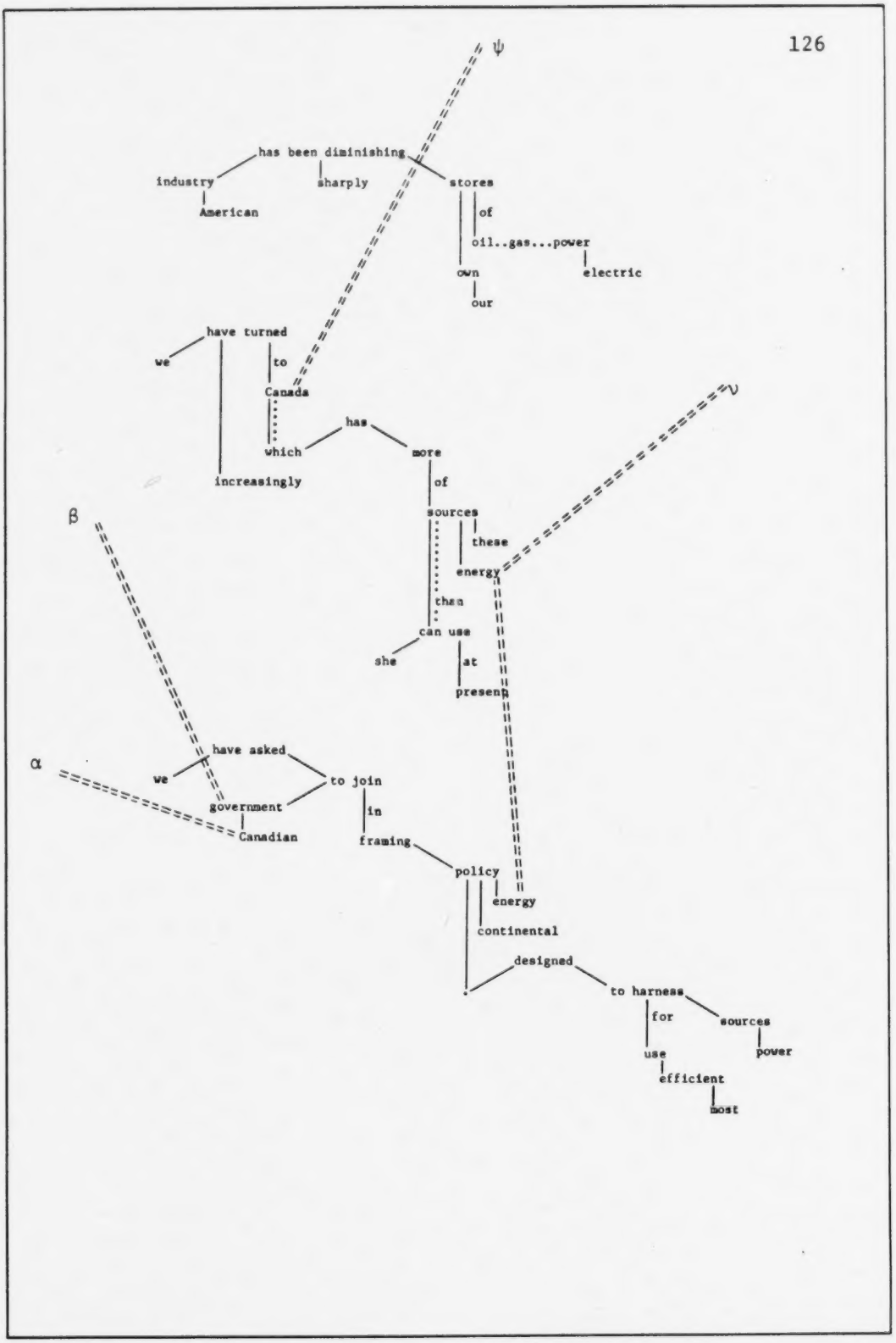

poses
power
problems
sheer
special
of
for
United States
Canada
own....contrbl
Americans
segment
vast
vital
that
of
such
economy
Canada's
feel
citizens
reduced
her
to
hewers.....drawers
of
of
wood
water

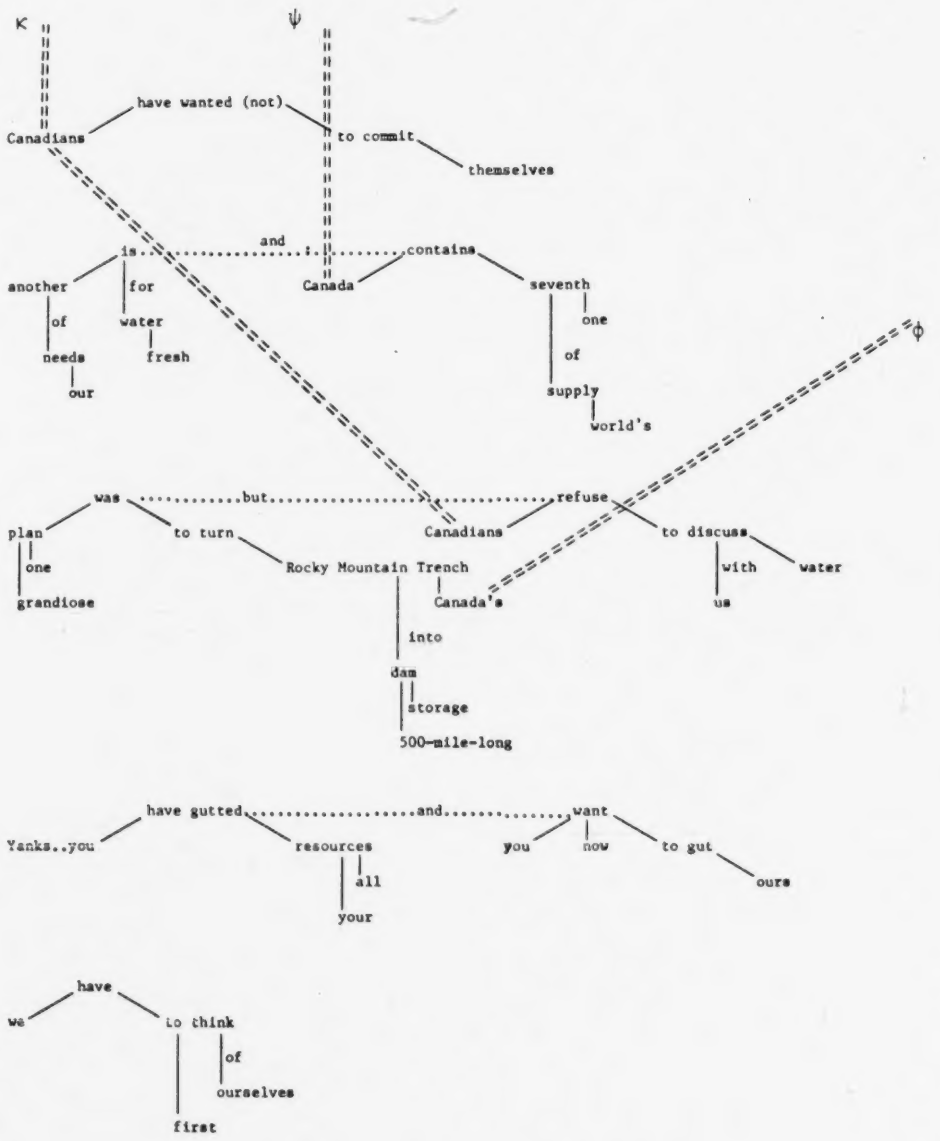

β ψ ϕ
have been moving
Canadians
toward
conclusion
that
are
they
subservient
to
United States
began
government
overhaul
Liberal
sweeping
Trudeau's
of
policy
foreign
was cut
commitment
in
troop
half
NATO
Canada's
in
Europe
recognized
and
signed
Ottawa
Red China
accord
with
Soviet Union
Claimed
Trudeau
on
that
needed
visit
Canada
contacts
to
because
world
Moscow
May
last
endangered
presence
identity
U.S.
her
overpowering
national
κ β ϕ ψ


α
ψ
is
issue
gut
extent
of
investment...ownership
U.S.
have doubled
investments
direct
in
Canada
since
1960
to
28 billion
are
sectors
richest
fastest-growing
of
economy
Canadian
foreign-controlled
have felt (never)
I
that
are
percentages
measure
true
of
independence
country's
any
have been
majority
of
subsidiaries
U.S.
vast
citizens
corporate
good
Canadian
ψ
λ
α

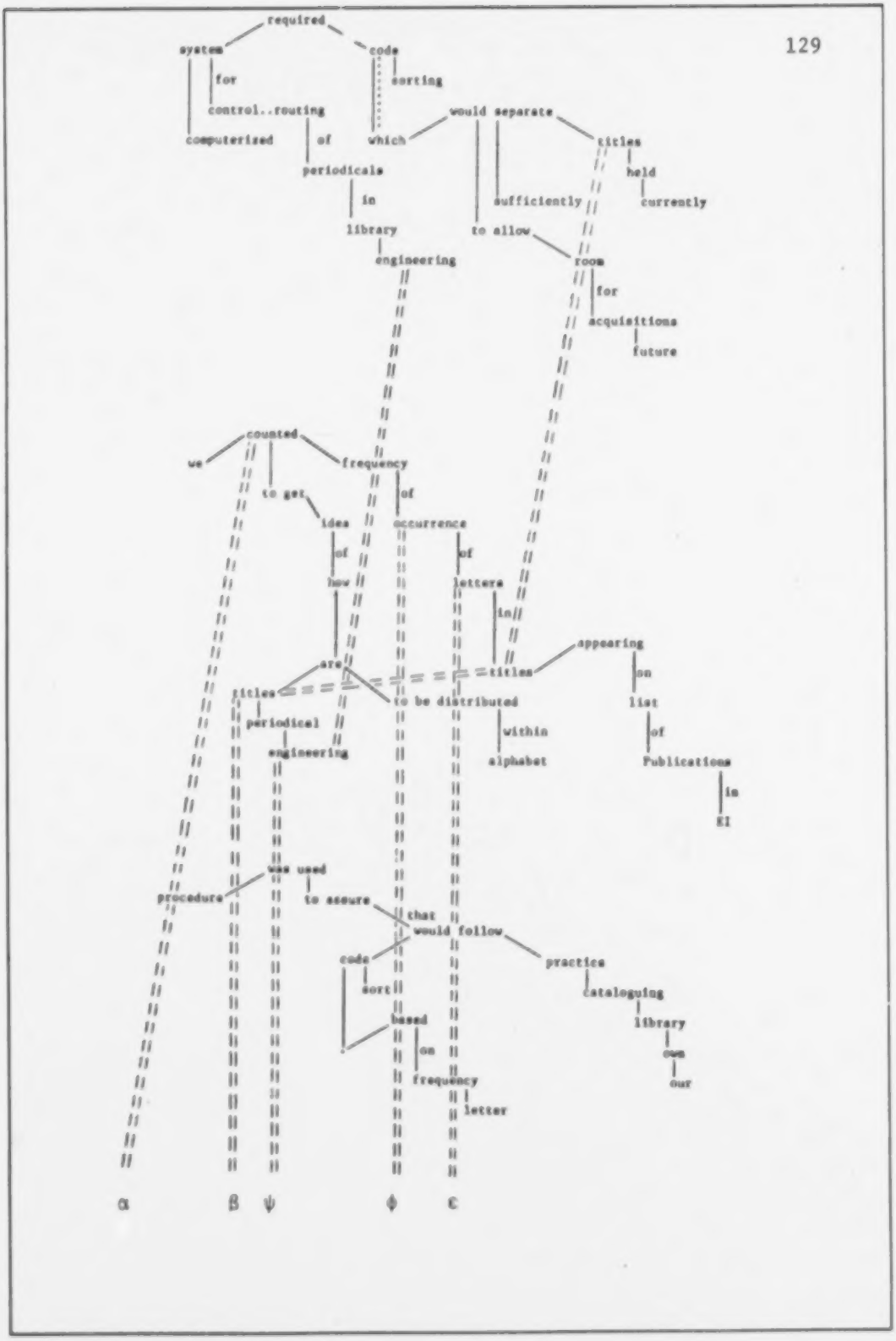

κ
λ
α
may have increased
it
sovereignty
rather than
diminished
Canadian
if
has boosted
and
remains
economy
yet
subject
to
laws ....taxes
her
money
U.S.
if
enable
it
to develop
Canadians
resources
while
at
home
following
policy
abroad
own
their
foreign
look
Canadians
upon
holding
U.S.
in
country
their
as
form
of
imperialism
economic
exploitive
κ
λ

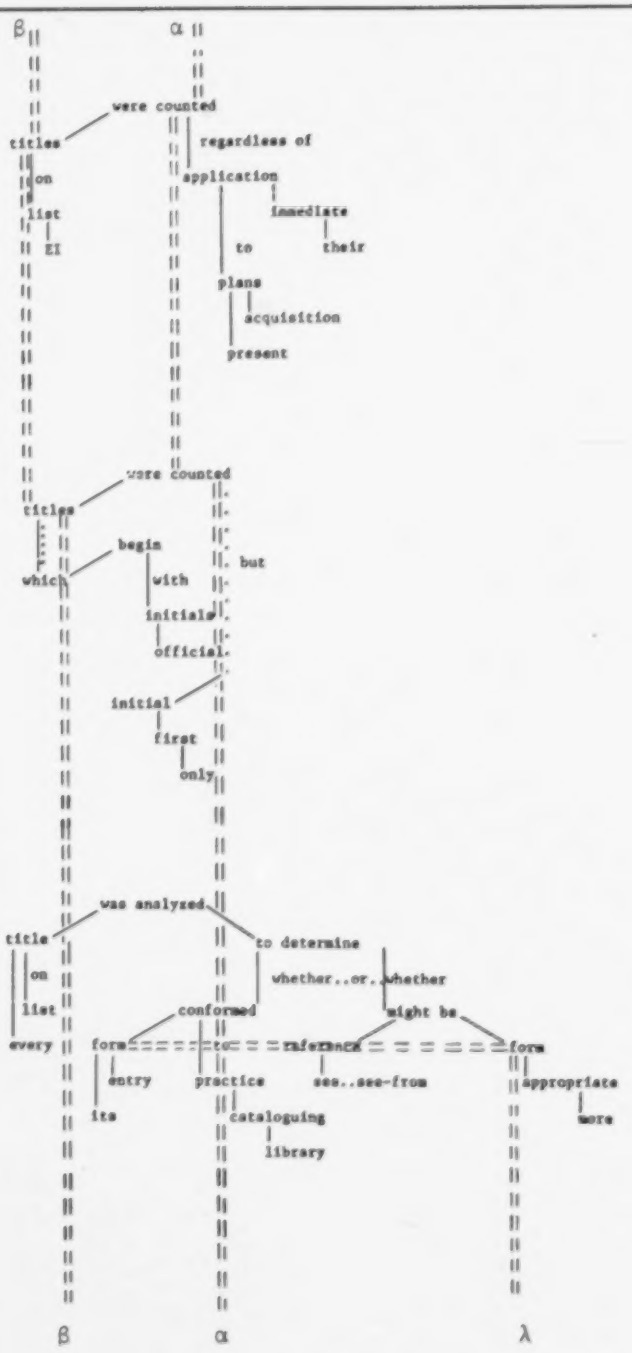

β
ψ
have joined
thousands
organization
lobby
nationwide
is pressuring
which
called
Committee
to buy
government
Canada
for
back
Canada
from
Independent
Americans
government
has created
Canada Development Corporation
with
million
$250
β
ψ
to begin
buying
into
sectors
key
of
economy
it
is expected
that
will set up
Ottawa
to monitor
oversee
agency
screening
inflow
of
capital
companies
foreign-owned
is
bound
question
as
to get
of
touchy
worse
resources
energy
ν

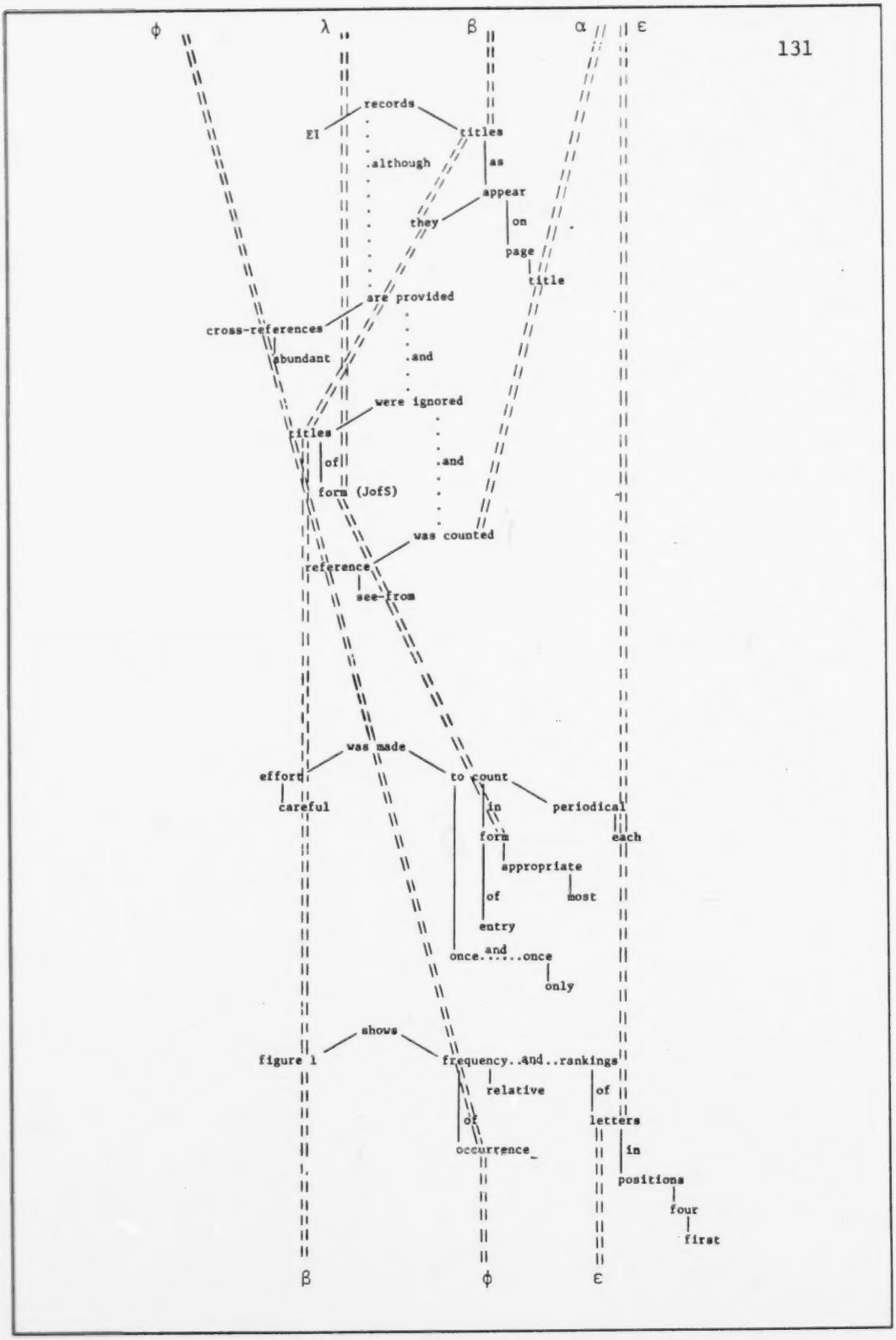

ψ
has been diminishing
industry
American
sharply
stores
of
oil..gas...power
electric
own
our
have turned
we
to
Canada
which
has
more
of
increasingly
sources
these
energy
than
can use
she
at
present
ν
β
have asked
we
to join
α
government
Canadian
in
framing
policy
energy
continental
designed
to harness
for
sources
power
use
efficient
most

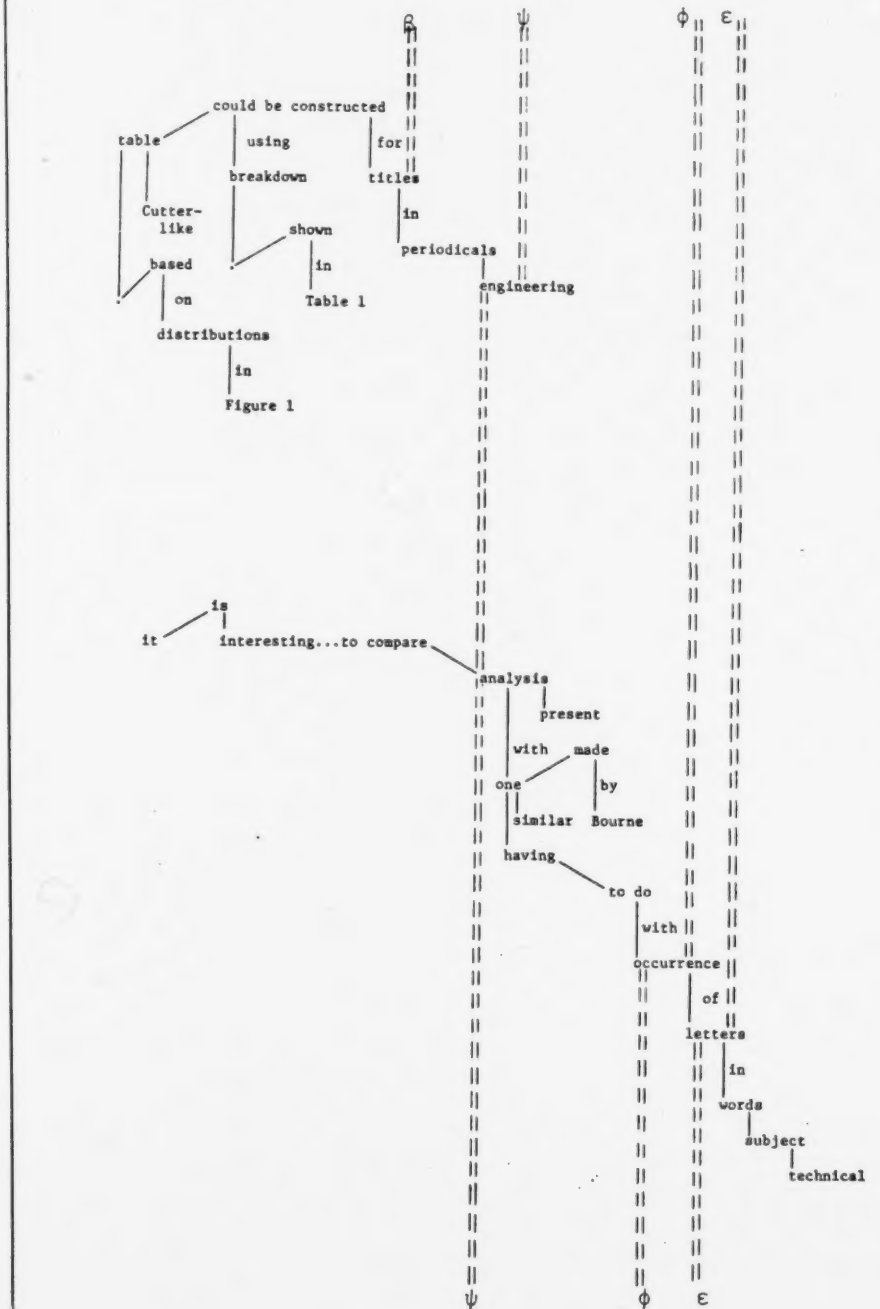

K
Ψ
Canadians
have wanted (not)
to commit
themselves
and
is
contains
another
of
needs
our
for
water
fresh
Canada
seventh
one
of
supply
world's
Φ
was
but
refuse
plan
one
grandiose
to turn
Rocky Mountain Trench
Canadians
to discuss
with
us
water
Canada's
into
dam
storage
500-mile-long
Yanks..you
have gutted
resources
all
your
and
want
you
now
to gut
ours
we
have
to think
of
ourselves
first

3. AGNES graph of a short technical article, "Occurrence of Letters in Engineering Periodical Titles" (d3).


required
system
code
for
sorting
control..routing
would separate
computerized
of
which
titles
held
periodicals
currently
in
sufficiently
library
to allow
engineering
room
for
acquisitions
future
counted
we
frequency
to get
of
idea
occurrence
of
of
how
letters
in
are
titles
appearing
titles
to be distributed
on
periodical
list
within
engineering
of
alphabet
Publications
in
EI
was used
procedure
to assure
that
would follow
code
practice
sort
cataloguing
based
library
on
own
frequency
our
letter
α
β
ψ
φ
ε


β
α
were counted
titles
regardless of
on
application
list
immediate
EI
to
their
plans
acquisition
present
were counted
titles
begin
but
which
with
initials
official
initial
first
only
was analyzed
to determine
title
on
whether..or..whether
list
conformed
might be
every
form
to
reference
form
entry
practice
see..see-from
appropriate
its
cataloguing
more
library
β
α
λ


φ
λ
β
α
ε
records
titles
El
.although
as
appear
they
on
page
title
are provided
cross-references
abundant
.and
were ignored
titles
of
form (JofS)
.and
was counted
reference
see-from
was made
effort
careful
to count
in
periodical
each
form
appropriate
most
of
entry
once.....and.....once
only
shows
figure 1
frequency..and..rankings
relative
of
of
occurrence
of
letters
in
positions
four
first
β
φ
ε


could be constructed
table
using
for
breakdown
titles
Cutter-
like
in
shown
periodicals
based
in
Table 1
engineering
on
distributions
in
Figure 1
is
it
interesting...to compare
analysis
present
with
made
one
by
similar
Bourne
having
to do
with
occurrence
of
letters
in
words
subject
technical

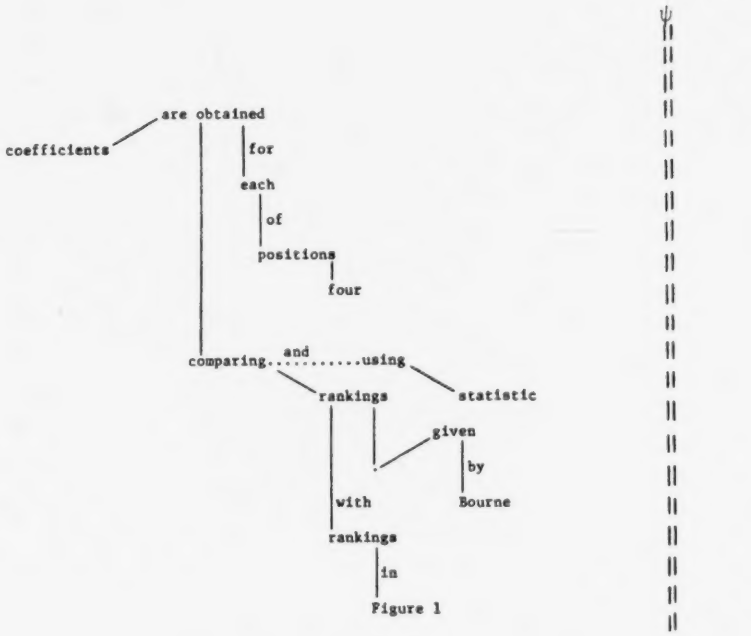
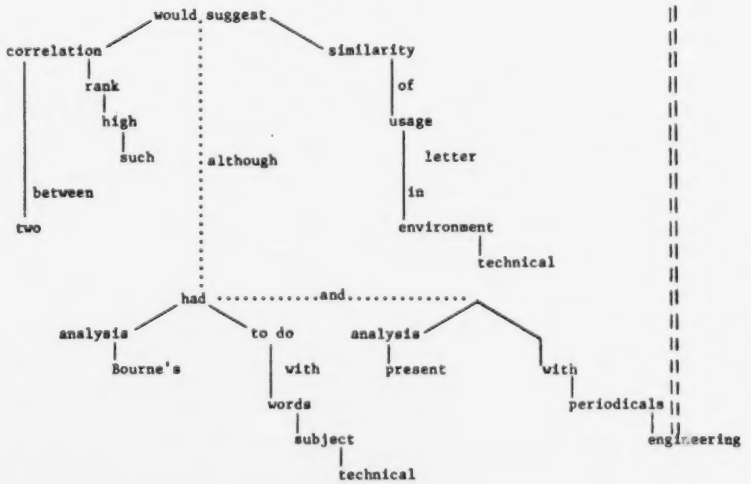

are obtained
coefficients
for
each
of
positions
four
comparing
and
using
rankings
statistic
given
by
Bourne
with
rankings
in

Figure 1


would suggest
correlation
rank
high
such
between
two
although
similarity
of
usage
letter
in
environment
technical

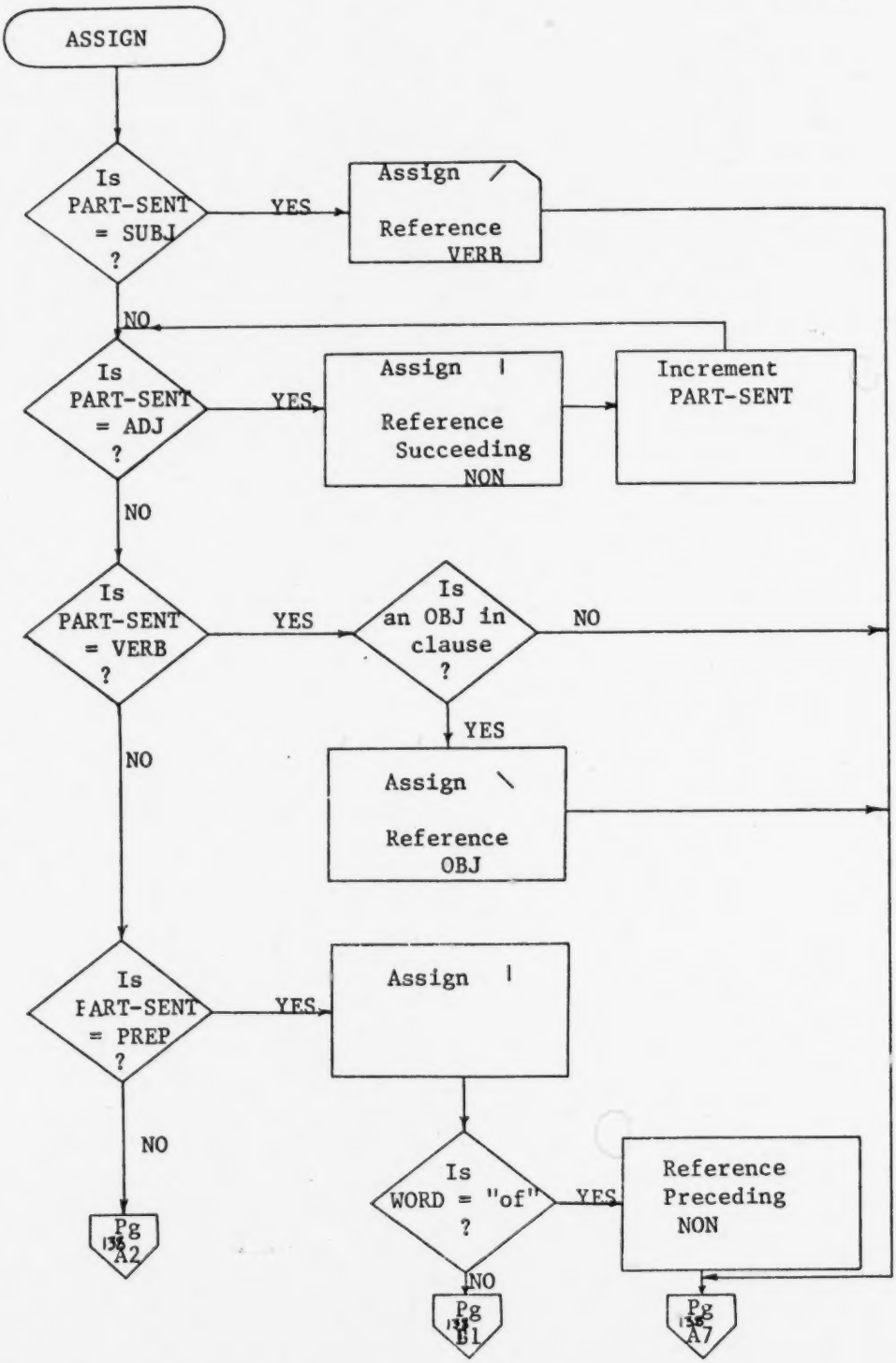
had
and
analysis
Bourne's
to do
with
words
subject
technical
analysis
present
with
periodicals
engineering

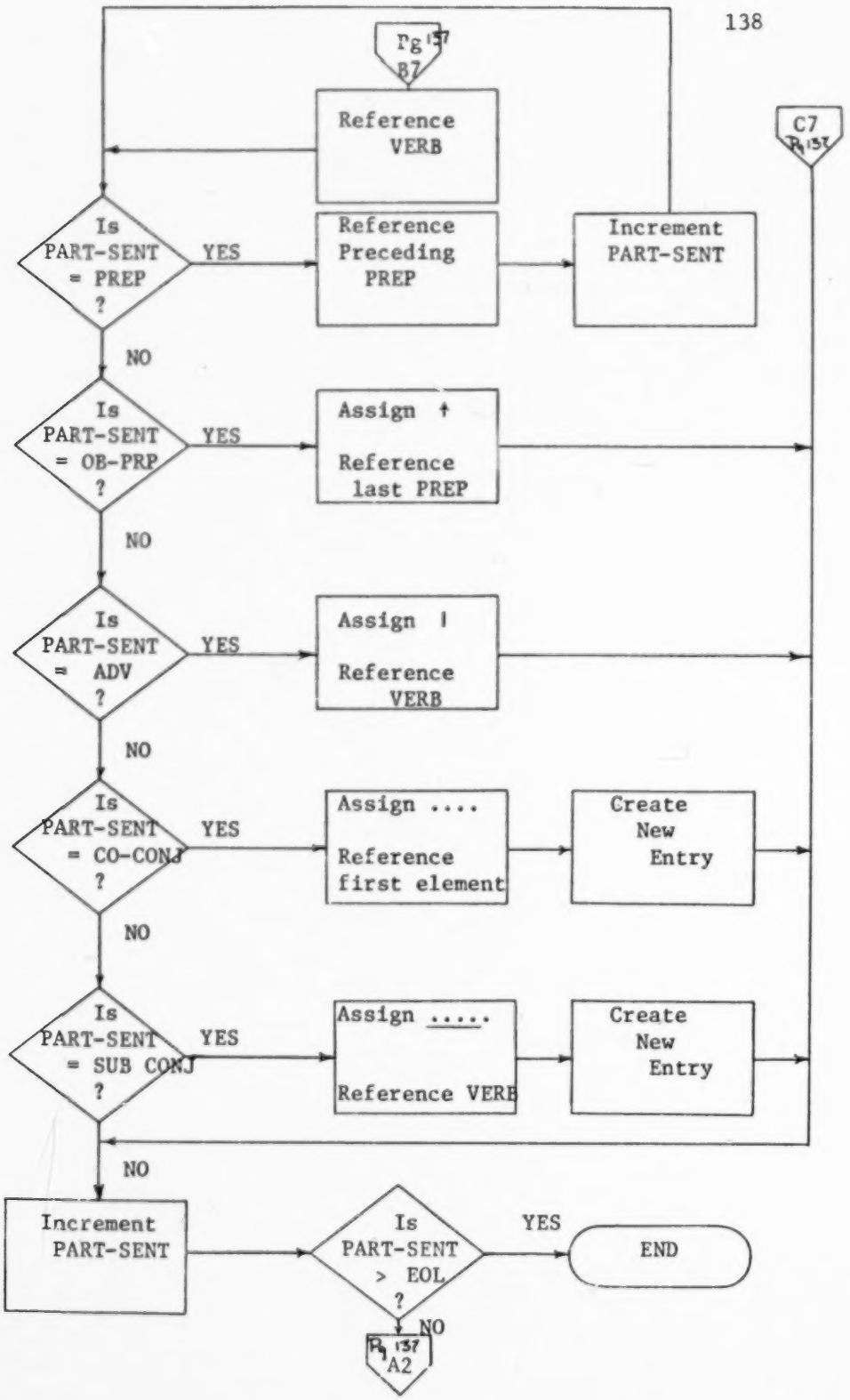
ϕ
ε
is
occurrence
similar
of
without
to
letters
regard
occurrence
in
to
fairly
of
position
position
letters
2-4
when
in
text
normal
compared
and
with
with
findings
which
in
of
Griffith
compares
study
Bourne
findings
based
his
on
own
words
107,000

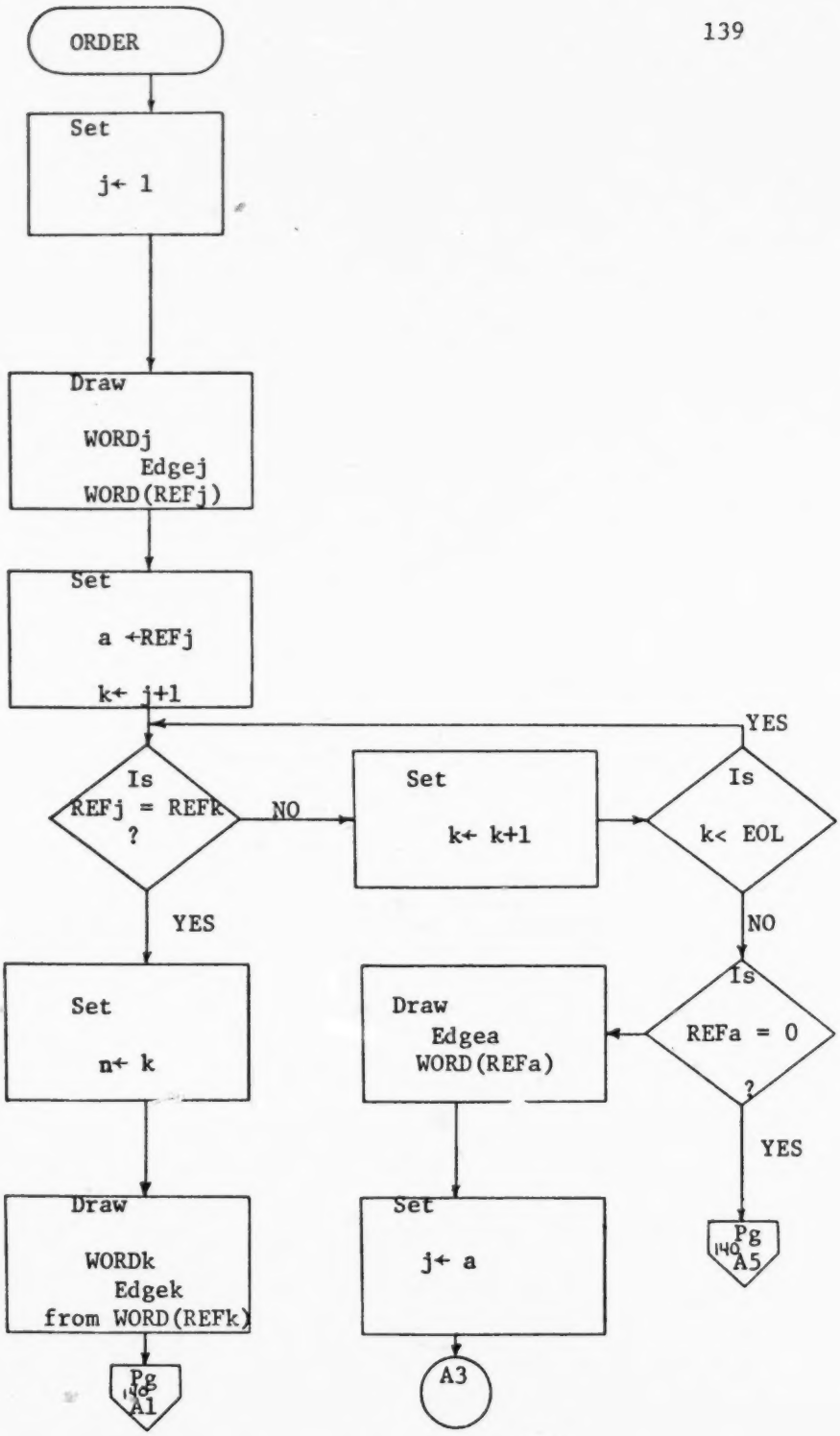
APPENDIX B

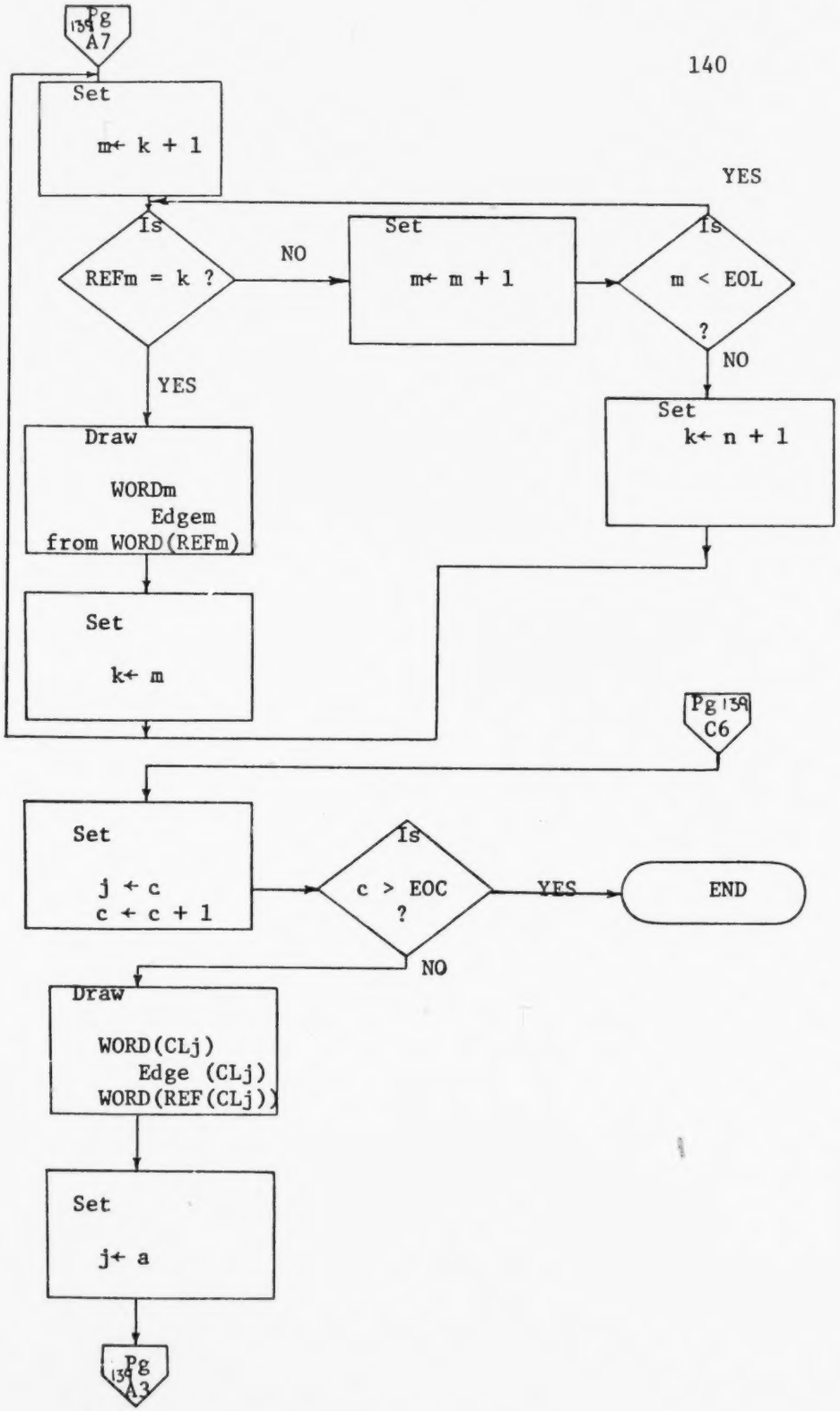
APPENDIX B: FLOWCHARTS FOR COMPUTERIZATION OF AGNES

The flowcharts which follow define the two major functions which must be performed in order to computerize the algorithm which has been suggested. The first routine ASSIGN assigns an appropriate edge and referent to each word in the input sentence on the basis of part of sentence assignments. The second routine ORDER orders the construction of the graphs from the table produced by ASSIGN.


ASSIGN
Is PART-SENT = SUBJ ?
YES
Assign Reference VERB
NO
Is PART-SENT = ADJ ?
YES
Assign Reference Succeeding NON
Increment PART-SENT
NO
Is PART-SENT = VERB ?
YES
Is an OBJ in clause ?
NO
YES
Assign Reference OBJ
NO
Is PART-SENT = PREP ?
YES
Assign
NO
Pg 135 A2
Is WORD = "of" ?
YES
Reference Preceding NON
NO
Pg 135 B1
Pg 135 A7


Pg 137 B7
Reference VERB
C7 Pg 137
Is PART-SENT = PREP ?
YES
Reference Preceding PREP
Increment PART-SENT
NO
Is PART-SENT = OB-PRP ?
YES
Assign ↑
Reference last PREP
NO
Is PART-SENT = ADV ?
YES
Assign ↓
Reference VERB
NO
Is PART-SENT = CO-CONJ ?
YES
Assign ....
Reference first element
Create New Entry
NO
Is PART-SENT = SUB CONJ ?
YES
Assign .....
Reference VERB
Create New Entry
NO
Increment PART-SENT
Is PART-SENT > EOL ?
YES
END
NO
Pg 137 A2


ORDER
Set
j← 1
Draw
WORDj
Edgej
WORD(REFj)
Set
a ←REFj
k← j+1
Is
REFj = REFk
?
NO
YES
Set
k← k+1
Is
k< EOL
YES
NO
Set
n← k
Draw
WORDk
Edgek
from WORD(REFk)
Pg
140
A1
Draw
Edgea
WORD(REFa)
Is
REFa = 0
?
YES
Set
j← a
A3
Pg
140
A5


Pg 139 A7
Set
m← k + 1
YES
Is
REFm = k ?
NO
Set
m← m + 1
Is
m < EOL
?
NO
YES
Set
k← n + 1
Draw
WORDm
Edgem
from WORD(REFm)
Set
k← m
Pg 139 C6
Set
j ← c
c ← c + 1
Is
c > EOC
?
YES
END
NO
Draw
WORD(CLj)
Edge (CLj)
WORD(REF(CLj))
Set
j← a
Pg 139 A3